# Hydrological ensemble prediction systems: from evaluating daily streamflow forecasts to exploring the impact of selected flood events in a future climate

Trine Jahr Hegdahl

Thesis submitted for the degree of Philosophiae Doctor (PhD)

Department of Geosciences
Faculty of Mathematics and Natural Sciences
University of Oslo
Oslo, Norway
2021

# Abstract

Floods can be devastating, and information about possible flood impact and early flood warnings are essential for public safety and response. Although weather forecasting is improving, weather predictions continue to be uncertain. Systematic errors originate from the difference between the scale of global forecasting models and the local scale needed for hydrological impact modeling. Ensemble prediction systems can capture uncertainty in the meteorological forecast variables, and different techniques exist to reduce systematic errors in both meteorological forecasts and the hydrological output. In a flood forecasting system, warnings are issued when a forecasted flood is likely to exceed a predefined flood threshold. In such systems, there is a balance between issuing too many false alarms and failing to detect flood events, the latter being considered worse.

Flood exceedance levels are used to define flood-prone areas and are essential for establishing guidelines for construction and placement of infrastructure and buildings. In the context of climate change, flood levels should be adapted and adjusted for future environments. However, the effect of climate change at the regional scale, important for local adaptation, is not necessarily fully reflected by the global model output. In some regions, both flood size and patterns will change and raise new challenges and demands for local communities. However, it may be difficult for the decision-maker to relate information about climate change; therefore, alternative approaches to demonstrating possible scenarios of future flood impact are useful.

This study had two focuses to address forecasting of floods and assessment of future flood impact.

(1) The first was to evaluate different pre- and postprocessing techniques applied to more than 100 catchments in Norway. The catchments represent the diversity of the Norwegian hydroclimate, ranging from humid coastal to subarctic continental. The evaluation was targeted at revealing any seasonal and regional differences. The study also assessed the differences in performance of the processing schemes for the full dataset, including all the

days of the study period, and a flood dataset, which only included days when streamflow exceeded the mean annual flood.

The results of the forecasting studies showed that the processing schemes improved model performance for nearly all catchments and lead times, and that there were regional differences in the performance of the pre- and postprocessing schemes under flood conditions. The combination of preprocessing and postprocessing performed best for inland and high-elevation catchments, whereas for coastal catchments, preprocessing precipitation using Bayesian model averaging was most effective for short lead times and preprocessing temperature was more important for longer lead times. Independent of processing scheme, spring floods were more predictable for longer lead times than autumn floods, which were not predictable beyond 3–4 days ahead.

(2) Atmospheric rivers are responsible for most large floods in the west coast of Norway. The second focus of this study was to assess the impact of these events in a future warmer climate. An event-based storyline approach was applied to explore future climate events and compare them to present climate events. The storyline approach enables high-resolution models better adapted to resolving the processes and orography responsible for the extreme precipitation causing flooding in western Norway. A modeling chain, similar to the operational weather and flood forecasting systems, was applied; it included a high-resolution global climate model, a non-hydrostatic weather forecasting model, and 37 hydrological catchment models. By applying a storyline approach, this study aimed to contribute to the present knowledge about regional flood impact in a future climate. The operational forecasting chain provides a flood warning setting that is known to end users; therefore, it may facilitate the interpretation of results, enabling reference to relevant events previously experienced.

The high-resolution models applied in this study are well adapted to simulating atmospheric river events making landfall in the west coast of Norway. By including a non-hydrostatic regional weather forecasting model, the distribution of precipitation intensities in the complex topography was further improved. The two most extreme events for 30 years of present and future climate were selected. The modeling results showed that both

future events would involve larger floods in more catchments compared to the present climate events. Although meteorological forcing is the most important variable for flood estimates, the initial hydrological conditions influenced the ultimate flood level and is, therefore, an additional aspect in the assessment of any plausible worse case. These combinations of factors can be another contribution to the storyline approach, in which the addition of relevant elements within limits of physical plausibility imitates stress testing commonly used in engineering and other disciplines.

# Acknowledgment

I will start by thanking main-supervisor Kolbjørn Engeland; always positive, encouraging, and available for questions and suggestions at almost any hour. You are a rock! Co-supervisor Lena Tallaksen I will thank for your clear and precise feedbacks, and for being encouraging when I returned to university applying for a new Master in 2012. Co-supervisors Ingelin Steinsland and Andrew Singleton I thank for providing valuable contributions and discussions throughout my PhD. I am lucky to have had a team of supervisors, and you have all contributed with your specific knowledge from different professional backgrounds and experiences. It has been a joy working with you all!

A special thanks to Jana Sillmann for including me in the TWEX project. Nathalie Schaller, Malte Müller, and all participants within the TWEX group, thank you for all new insights and interesting discussions on storylines. At MET Norway I will thank T. Nipen and I. Seierstad for their support and sharing of ensemble forecast knowledge. At ECMWF a special thanks to Linus Magnusson and Fredrik Wetterhall for providing the experimental dataset used to compare different ensemble forecasting cycles, and Fredrik for organizing my visiting stay at ECMWF.

To my colleagues at the hydrological modeling section at NVE, thank you for making work so interesting and providing such a great learning-environment. Bård Grønbech, thank you for the help in setting up the hydrological model for ensemble forecasting, and not least technical support when models won't behave! There are so many great colleagues at NVE, thank you all for contributing to a working environment that is positive, inspirational, including, and supportive. And last but not least, Hege Hisdal for giving me the opportunity to realize this PhD. Thank you so much!

And finally, my family; Tuva, Olve, and Yngve. Love you for being who you are, keeping me grounded, and making life complete!

# Contents

# List of papers

**Paper I**
**Hegdahl, T. J.**, Engeland, K., Steinsland, I., and Tallaksen, L. M., 2019: Streamflow forecast sensitivity to air temperature forecast calibration for 139 Norwegian catchments, *Hydrology and Earth System Sciences*, 23, 723–739, https://doi.org/10.5194/hess-23-723-2019.

**Paper II**
**Hegdahl, T. J**, Engeland, K, Singleton, A, and Steinsland, I., 2021: The benefits of pre- and postprocessing streamflow forecasts for an operational flood-forecasting system of 119 Norwegian catchments, *Hydrol. Earth Syst. Sci. Discuss. [preprint],* https://doi.org/10.5194/hess-2021-13, in review, 2021.

**Paper III**
**Hegdahl, T. J**., Engeland, K., Müller, M., and Sillmann, J., 2020: An Event-Based Approach to Explore Selected Present and Future Atmospheric River–Induced Floods in Western Norway. *Journal of. Hydrometeorology*, 21, 2003–2021, https://doi.org/10.1175/JHM-D-19-0071.1.

**Paper IV**
Schaller, N., Sillmann, J., Müller, M., Haarsma, R., Hazeleger, W., **Hegdahl, T. J**., Kelder, T., van den Oord, G., Weerts, A., & Whan, K., 2020: The role of spatial and temporal model resolution in a flood event storyline approach in western Norway. *Weather and Climate Extremes*, 29, [100259]. https://doi.org/10.1016/j.wace.2020.100259

For Papers I, II, and III, I was responsible for programming, hydrological modeling, analysis, and writing the manuscripts. All co-authors helped reviewing the Papers. Data from the meteorological models ECMWF IFS, EC-Earth, and AROME-MetCoOp were provided by ECMWF, KNMI, and MetNorway respectively. In Paper III, M. Müller provided the analysis and drafted the section of EC-Earth and AROME-MetCoOp. For Paper IV, N. Schaller was responsible for the main analysis and in writing the manuscript. I was responsible for the programming, modeling, analysis, and writing the section on hydrological modeling using HBV lumped. I also contributed reviewing Paper IV.

x

# Symbols and abbreviations

AR            Atmospheric river

AROME-MetCoOp            Applications of Research to Operations at Mesoscale - Meteorological Cooperation on Operational Numerical Weather Prediction

BMA            Bayesian model averaging

CAL            Internal notation for calibration by methods other than BMA

CMIP5            Coupled Model Intercomparison Project (Taylor et al. 2012)

CRPS            Continuous ranked probability score

EC-Earth            European Community Earth-System Model

ECMWF            European Centre for Medium-Range Weather Forecasts

EMOS            Ensemble model output statistics

ENS            Ensemble prediction system of ECMWF, earlier referred to as EPS

GCM            Global climate model

HBV            Hydrologiska Byråns Vattenbalansavdelning

HRes            High resolution deterministic forecast from ECMWF (model/model resolution different from ENS)

IFS            Integrated forecasting system of ECMWF

IPCC            Intergovernmental Panel on Climate Change

NSE            Nash–Sutcliffe efficiency

| | |
|---|---|
| RCM | Regional climate model |
| RCP | Representative concentration pathway |
| SeNorge | Interpolated observations at 1-km grid resolution for all of Norway |
| SPPT | Stochastically perturbed parameterization tendency (ECMWF model perturbation) |

| | |
|---|---|
| $P_o$ | Reference precipitation (mm) (SeNorge) |
| $P_{bma}$ | BMA preprocessed precipitation ensemble forecasts (mm) |
| $P_{cal}$ | CAL preprocessed precipitation ensemble forecasts (mm) |
| $P_{ens}/P_{raw}$ | Raw unprocessed precipitation ensemble forecasts (mm) |
| $T_o$ | Reference temperature (°C) (SeNorge) |
| $T_{bma}$ | BMA preprocessed temperature ensemble forecasts (°C) |
| $T_{cal}$ | CAL preprocessed temperature ensemble forecasts (°C) |
| $T_{ens}/T_{raw}$ | Raw unprocessed temperature ensemble forecasts (°C) |
| $Q_o$ | Reference streamflow (m$^3$/s) |
| $Q_{raw}$ | Raw (not postprocessed) ensemble streamflow (m$^3$/s) |
| $Q_{bma}$ | Postprocessed ensemble streamflow (m$^3$/s) |

# Chapter 1
# Introduction

## 1.1 Motivation

Floods contribute to about a third of the global economic losses due to natural disasters (Berz 2000; Jacobeit 2003). Flood warning systems are used for immediate action to save lives and reduce social and economic losses, whereas from a climate change perspective, knowledge of future floods is essential for planning to reduce the susceptibility of communities to floods.

In Norway, the first known flood warning was issued on April 6, 1860, almost two months before the extreme flood of that year (Roald 2013). This first warning was based on the knowledge that large amounts of snow in the mountains increased the probability of spring floods caused by snowmelt. Years later in 1938, flood information was shared by establishing communication lines for stations upstream to inform those downstream about the flood situation. In 1967, following the disastrous floods of the previous year, the first flood forecasting service for selected municipalities in the Glomma catchment was initiated. Regression equations based on upstream measurements were used to estimate the downstream flood levels (Hegge 1968). However, it took until the large flood of 1987 for a national flood forecasting center to be established in 1989. In 1995, yet another large flood caused severe damage in the Glomma and Lågen catchments, and the events further strengthened the flood forecasting service. By November 1995, precipitation-runoff models, using weather forecasts as forcing, were operational for 17 catchments. However, the computational time was long, and hence the usefulness of the model output was limited

(NOU-16 1996). The current operational flood warning system has its origin in the 1995 system, and with updates and developments now includes 145 catchments and three hydrological models run at different temporal and spatial resolutions.

Rainfall is the single most important variable causing floods; however, in cold climates, snow plays an additional role. In Norway, snowmelt-induced floods are typical of spring and early summer. Rain-induced floods can be linked to specific weather systems. The rare southeastern storm trajectories have been responsible for the largest known floods in eastern Norway, whereas in southern and central Norway, the southwestern storm trajectories are the most frequent flood generating systems (Roald 2008, 2013). The atmospheric river (AR) is a special variation of the western weather system, and is defined as a narrow filament of very high atmospheric moisture content that is transported from the tropical or extratropical latitudes toward the poles (e.g., Zhu and Newell 1998; Ralph and Dettinger 2011; Ralph et al. 2017). Atmospheric rivers cause intense precipitation when orographically lifted reaching a topographic barrier like the western mountain range of Norway (Stohl et al. 2008), and have been responsible for the largest floods in the region (Stohl et al. 2008; Lavers and Villarnini 2013, 2015; Azad and Sorteberg 2017; Benedict et al. 2019). While AR-related floods are characteristic of autumn and winter in western coastal catchments (Azad and Sorteberg 2017), snowmelt-induced or combined snowmelt- and rain-induced floods are common in spring and early summer and are typical floods of inland and high-elevation areas (Roald 2013).

### 1.1.1 Flood forecasting

To provide the exact level of flooding is difficult, and in risk assessment and subsequent decision-making probabilities on flood exceedance are essential. Therefore, implementation of uncertainty quantification in a flood forecasting system sets an added value. Different approaches are available for assessing uncertainty. For example, the 1999 Norwegian warning system applied a combination of deterministic meteorological forecasts and hydrological and statistical models to quantify the total uncertainty in the hydrological forecasts. In the first step, statistical models (Follestad and Høst 1998) were applied to the deterministic temperature and precipitation forecasts to account for

uncertainty in the meteorological input, which created an ensemble of new values used as input for hydrological modeling. In the second step, an autoregressive model describing the hydrological model error was used to give the total empirical distribution of composite errors forming the probability distribution of the streamflow (Langsrud et al. 1998, 1999). However, the methods were computationally intensive and required recalibration for each important update in the weather forecasting model. Another way of assessing the uncertainty that originates from meteorological forecasts is to use the meteorological ensemble forecast. Ensembles are typically produced by running the weather forecasting model with perturbations (small changes) of the initial conditions, model description, and parameterization, and thereby provide different trajectories of future states. Figure 1 illustrates the evolvement over time of different members of an ensemble, in which the distribution at the forecast time indicates the likelihood of occurrence.
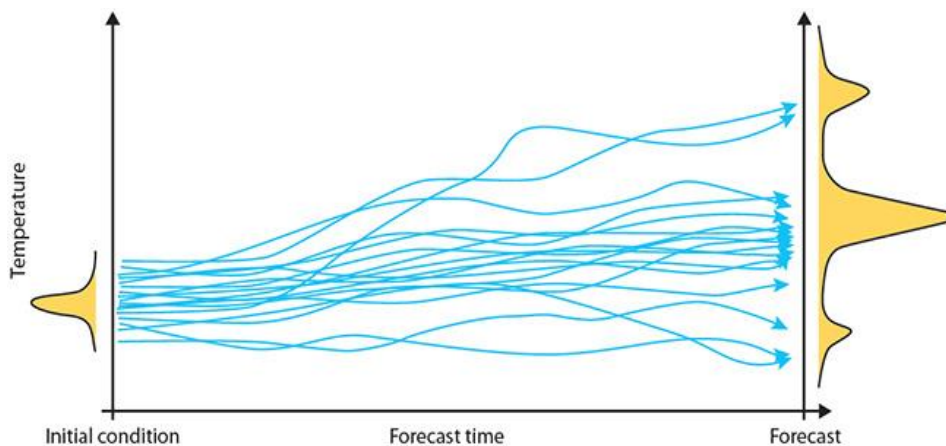


***Figure 1***. *An ensemble of temperature forecasts showing the range of possible values over the forecast time (ECMWF 2017).*

Although the skill of weather forecasting has improved in the last decades (e.g., Buizza 2005; Haiden et al. 2018), ensemble forecasts at the scale essential for hydrological modelling still contain errors in the mean and spread, i.e., they are biased and underdispersed (e.g., Buizza 1997; Wilks and Hamill 2007; Gneiting et al. 2007). A variety of techniques have been proposed to correct and improve the statistics of ensemble forecasts to provide calibrated and, thus, skillful forecasts. For an overview of techniques, see, for example, Vannitsem et al. (2018) and Li et al. (2017). A forecast is considered reliable when the observations appear to belong to the same probability distribution as the

forecast, e.g., for 90% of the forecasts, the observations are within a 90% confidence interval (Leutbecher and Palmer 2008).

When meteorological forecasts are used as input in hydrological models, the errors propagate through the model. Since the relationship between the input and the output from the hydrological model is nonlinear, it is essential to evaluate at what stage in the processing chain a correction scheme is most beneficial. In this thesis, the correction approaches applied to the meteorological inputs are referred to as preprocessing and those applied to the hydrological output as postprocessing.

Previous studies have analyzed the effects of both pre- and postprocessing on short- to medium-range streamflow forecasts (e.g., Zalachori et al., 2012; Roulin and Vannitsem, 2015; Benninga et al., 2017; Sharma et al., 2018). However, the conclusions on the ultimate effect of pre- and postprocessing are not consistent. Only a few studies have considered the effect on flood forecasts (e.g., Roulin and Vannitsem 2015; Fundel and Zappa et al. 2011). Therefore, there is a gap in the knowledge of the combined effect of pre- and postprocessing for a variety of catchments, and especially for floods.

### 1.1.2 Future flood predictions

The first assessment of changes in runoff and peak flows in Norway under a future warmer climate suggested that large snowmelt floods were likely to become rarer, whereas local rain-induced floods could become more severe due to increased intensity of rainfall (Sælthun et al. 1990). These early expectations were confirmed in later studies (Sælthun et al. 1998; Beldring et al. 2006; Roald et al. 2006; Lawrence and Hisdal 2011; Hanssen-Bauer et al. 2017). Climate change studies in Norway indicate regional differences in floods. The snowmelt floods typical of inland, northern, and eastern catchments would be reduced, and the seasonality of floods would change due to changes in temperature and precipitation. In a warmer climate, snowmelt floods would occur earlier and diminish in some regions (Vormoor et al. 2015). The west coast is most prone to increased precipitation and rain-induced floods. In the warmer future environment, the intensity and frequency of weather systems like atmospheric rivers are expected to increase (Whan et al. 2020), resulting in an increase in atmospheric river-induced floods. Although atmospheric rivers

and their future changes have received attention in recent climate studies (e.g., Dettinger 2011; Ralph and Dettinger 2011; Espinoza et al. 2018), the effect of atmospheric rivers on floods at the catchment scale is not well studied.

The general approach in climate projection studies is to apply a multimodel ensemble (e.g., Sillmann et al. 2013; Hanssen-Bauer et al. 2017), and thereby represent the uncertainty not only in the description of atmospheric processes, but also in the description, coupling, and interactions between the atmosphere, ocean, and land surface processes/schemes. However, there are situations where a different approach, referred to as a storyline, seems appropriate (IPCC 2010). For example, specific weather processes may be better described and resolved by one global climate model (GCM) than by others, and in such cases, the model best suited to solving the specific task should be selected (IPCC 2010). Application of storyline approaches is advancing to provide complementary information to probabilistic assessments. Focusing on a few specific events allows for high-resolution modeling that can resolve specific atmospheric processes not possible within a multimodel ensemble running at a coarse spatial-temporal resolution (Hazeleger et al. 2015; Shepherd et al. 2018; Shepherd 2019).

## 1.2 Objectives

The main objectives of this thesis were to explore how the hydrological ensemble modeling chain can be set up to achieve reliable flood forecasts for a variety of catchments in Norway, and to further assess the impact of specific extreme flood events in a future warmer climate.

Therefore, this study had two main focuses.

(1) It evaluated different pre- and postprocessing techniques applied to more than 100 catchments in Norway. All catchments are part of the Norwegian flood forecasting system and represent the diversity of the country's hydroclimatic conditions, ranging from humid coastal to subarctic continental. The evaluation was targeted to reveal seasonal and regional differences, in addition to the differences between forecasting floods and day-to-day forecasting.

The following research questions were answered:

*Q1: Will pre- and/or postprocessing schemes improve daily streamflow forecasts, and to which variables should a processing approach be applied?*

*Q2: How do the processing schemes perform under flood conditions, and are there seasonal or regional differences?*

(2) The second focus was on AR-induced floods in a future warmer climate. An event-based storyline approach was applied to explore and compare the future climate to the present atmospheric river events affecting western Norway. All analyses were set within a modeling chain similar to the operational weather and flood forecasting systems. The modeling chain included a high-resolution GCM, a nonhydrostatic weather forecasting model, and a hydrological model set up for 37 catchments. The thesis addressed two additional questions:

*Q3: What is the added value of an event-based storyline approach in climate impact studies?*

*Q4: What is the added value of higher spatial resolution in climate impact studies?*

## 1.3  Study design

This thesis consists of four papers; Paper I, III and IV are published in peer-reviewed journals, and Paper II is ready for submission. The operational modeling chain for flood forecasting was used as a basis to address the research questions. Figure 3 shows the three modeling chains used in the thesis, and Figure 4 gives an overview of the questions addressed and the approaches used in each paper. Papers I and II answered questions Q1 and Q2. Paper I addressed the effect of temperature calibration on streamflow forecasts and used modeling chain 1 to establish retrospective forecasts and modeling chain 2 to establish initial conditions for the forecasts and reference the streamflow used for evaluation of forecasts. In Paper II, the analysis was extended by addressing the combined effect of preprocessing temperature and precipitation and postprocessing streamflow on both streamflow and flood forecasts. Paper III addressed question Q3 and used modeling chain 3 to establish the flood scenarios and modeling chain 2 to establish the initial conditions. In this paper, the largest atmospheric river events in western Norway in the

present and future climate were compared using an event-based storyline approach. Paper IV addressed question Q4 and used modeling chain 3 to evaluate the added value of applying a nonhydrostatic weather forecasting model in the modeling chain.
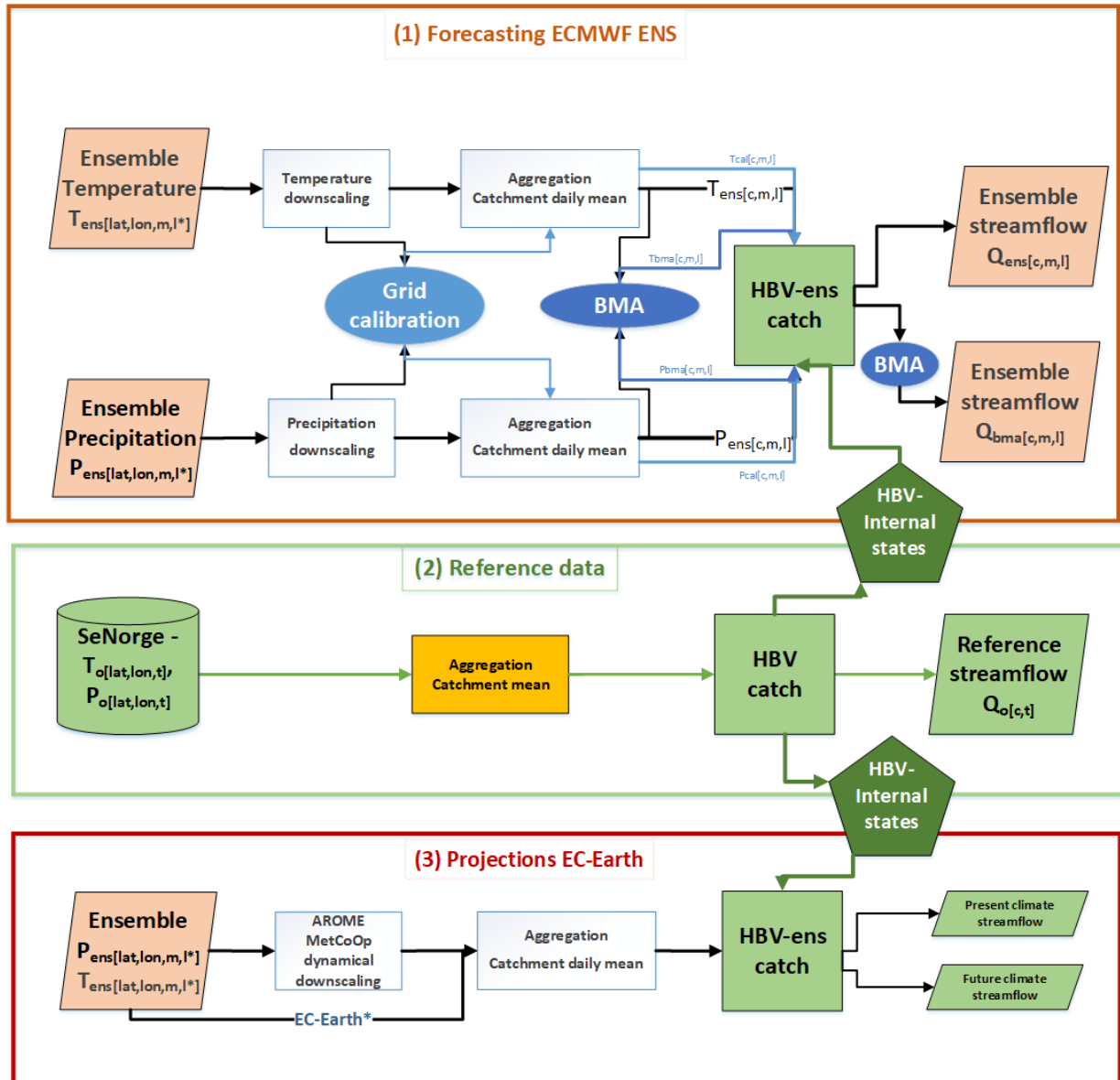


**_Figure 2._** _Overview of the three modeling chains used in the studies presented in this thesis. The same hydrological models (HBV-ens) were used for all chains. Modeling chain 1 was used to establish streamflow forecasts in which ECMWF ENS (European Center for Medium-Range Weather Forecasts - Ensemble prediction system) was used as meteorological forecasts and the internal states were obtained from modeling chain 2. In modeling chain 2, the seNorge (interpolated observations at 1-km grid resolution for all of Norway) daily catchment mean temperature (T) and precipitation (P) were used as a reference for the BMA (Bayesian model averaging) preprocessing, verification of the forecasts, and to establish the reference streamflow and the internal states of the hydrological model. The reference streamflow was used as a reference for postprocessing, and as verification for streamflow forecasts in modeling chain 1. In_

modeling chain 3, the downscaled AROME-MetCoOp (Applications of Research to Operations at Mesoscale - Meteorological Cooperation on Operational Numerical Weather Prediction) was primarily used as meteorological forcing, but EC-Earth (European Community Earth-System Model) was used directly to evaluate the effect of resolution of high impact studies. The initial hydrological conditions were established by the reference HBV in modeling chain 2. All modeling chains are explained in more detail in sections 3 and 4.
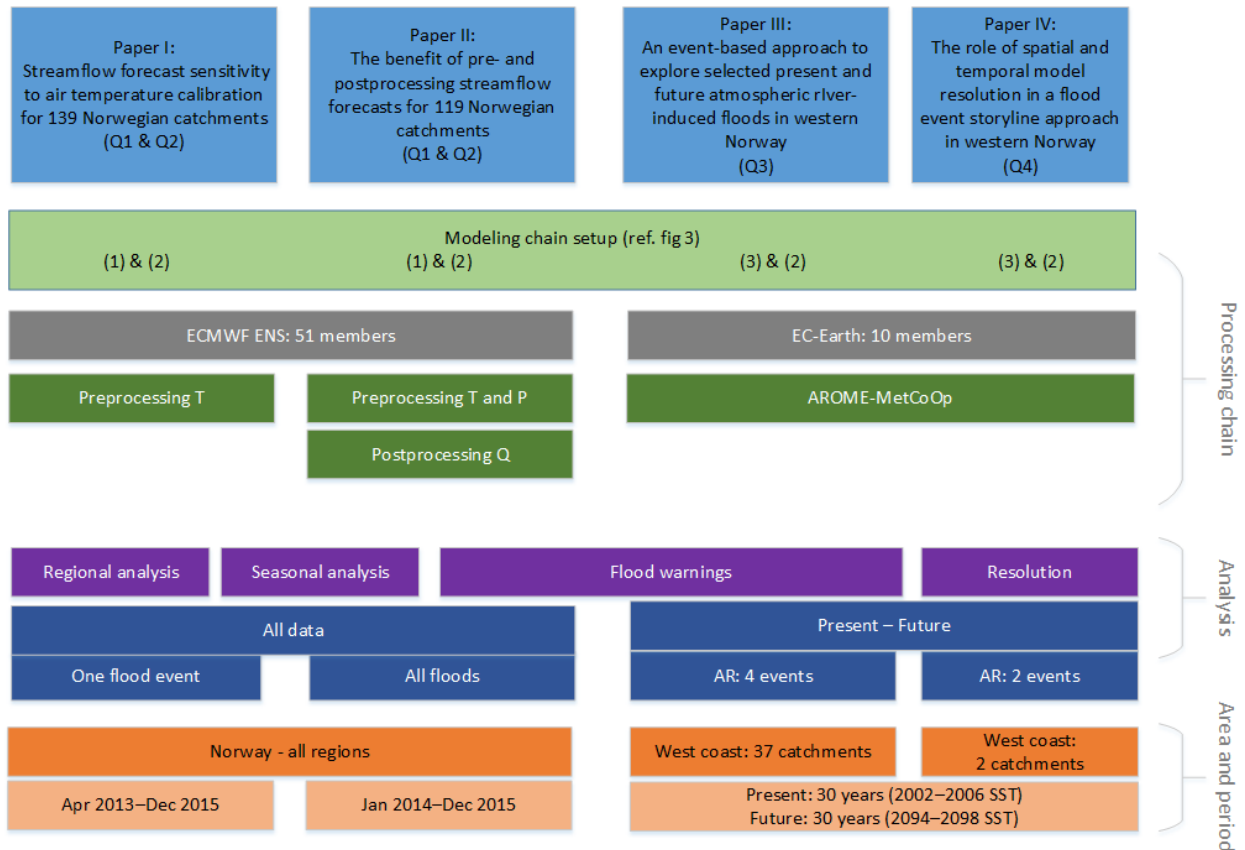


***Figure 3****. Overview of the four papers, including the research questions (Q), the modeling chain, analysis approaches, and underlying data used in each paper. Meteorological data by ECMWF ENS (European Center for Medium-Range Weather Forecasts - Ensemble prediction system), EC-Earth (European Community Earth-System Model), and AROME-MetCoOp (Applications of Research to Operations at Mesoscale - Meteorological Cooperation on Operational Numerical Weather Prediction). Input data used were temperature (T) and precipitation (P), whereas sea surface temperature (SST) defined the periods used to establish climate projection data. Extreme atmospheric river (AR) events were the basis for future flood evaluation.*

# Chapter 2
# Scientific background

Dynamical weather and climate models describe complex processes of the earth system and consist of differential equations that describe the motion of the atmosphere and conservation of energy and mass. Calculations are applied to grid cells in vertical layers covering the globe (Fig. 4); the resolution (size) of the grid cells determines which processes are resolved, which are parameterized, and how well the model represents the topography. The hydrological cycle involving storage and transport of water in the atmosphere, ocean, cryosphere, and land surface (Fig. 4, left) is an important contribution to the energy exchange between the land surface and the atmosphere, as well as for the spatial distribution of water. Surface water and rivers are quantitatively a small part of the hydrological cycle, but are a vital part of livelihoods and location of settlements.

The focus of this study was to improve estimations of streamflow, which included both enhancing the flood forecasts up to a lead time of nine days, and to explore specific flood events in a future warmer climate. For both daily streamflow forecasts and flood projections in a future climate, the weather variables (e.g., temperature and precipitation) from global weather or climate models were used as input for hydrological models. The meteorological variables are often provided at a coarser spatial scale compared to the local scale of the hydrological model. To achieve meteorological variables that represents the scale of the hydrological model and correspond well with observations at the land surface, a processing step that often included both downscaling and calibration (e.g., bias correction) was applied to the raw meteorological forecasts.
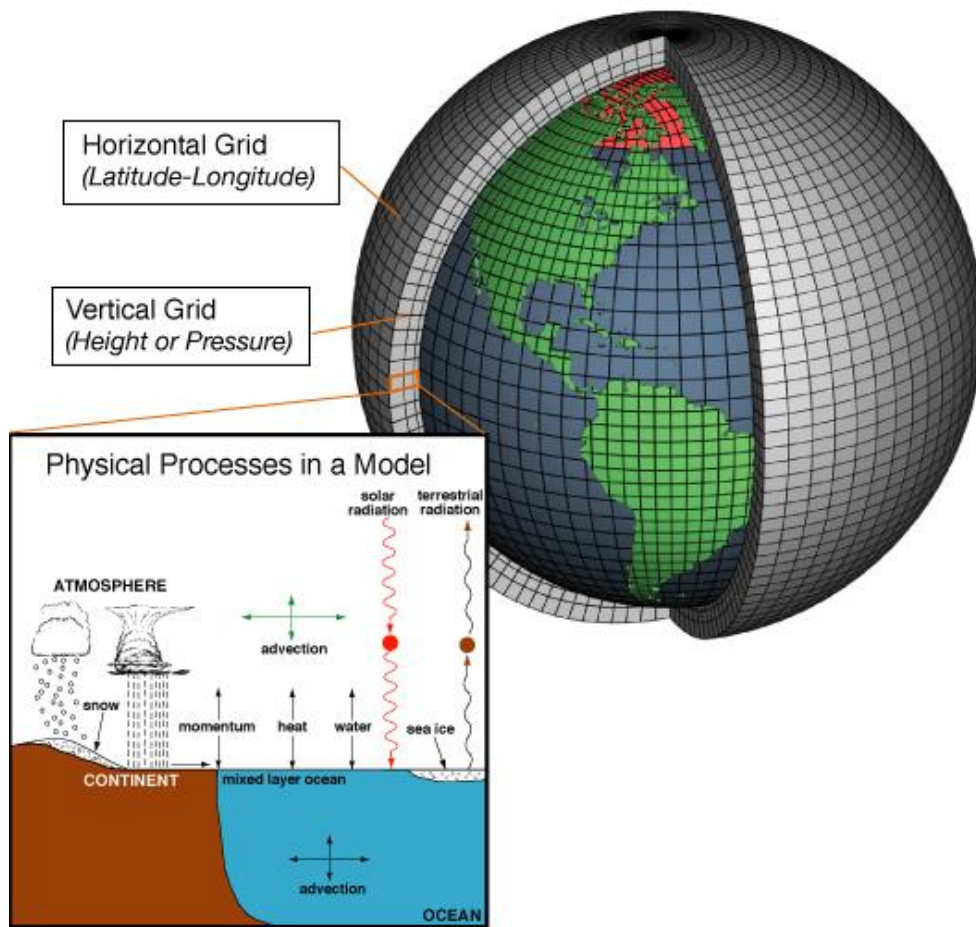
**Figure 4.** *Simplified view of a dynamic weather and climate model in which the planet is divided into a 3-dimensional grid, and within each grid, wind, heat transfer, radiation, relative humidity, and surface hydrology are calculated and interactions with neighboring grids are evaluated (https://celebrating200years.noaa.gov/breakthroughs/climate_model/modeling_schematic.html).*

## 2.1  Hydrological ensemble forecasts

Hydrological forecasts are subject to uncertainty from various sources, including meteorological forcing, initial conditions, hydrological model parameters, and model structure. The uncertainties might be both random, almost independent between time steps, or systematic, like systematic biases that persist over longer time windows (Engeland et al. 2016). In hydrological forecasting, both interpolated observations and forecasts are used as forcing. For observations there are errors in measurements. Observed precipitation often suffers from systematic errors/uncertainty due to under-catch that depends on wind speed and precipitation phase (Wolff et al. 2015). In addition, temperature and precipitation need to be interpolated from the station points to the catchment area used in the hydrological

model. The interpolation introduces both random and systematic errors where the latter is caused by the representativeness of the station location, for example how well they represent the elevation distribution of the catchment. The interpolated forcing is used to establish the initial conditions of the hydrological model before forecasts are used. Uncertainties in the meteorological observations therefor translates into the initial hydrological conditions. To quantify uncertainties in meteorological forecasts, most weather forecasting centers use ensemble forecasts, which are increasingly being used also by hydrological forecasting centers (Cloke and Pappenberger 2009; Wetterhall et al. 2013). Uncertainties is the meteorological forecasts are described in section 2.2. Streamflow observational uncertainty can be caused by stage measurement or by the rating curves, where floods and high streamflow values can be more uncertain (e.g., Reitan and Petersen-Øverleir 2009). The hydrological models are simplifications of the natural process, and therefore uncertainty due to the model structure can be addressed by a multimodel approach. For example, Thiboult et al. (2016) improved the streamflow forecasts throughout the forecast horizon up to nine days ahead by including multiple models. The parameter uncertainty might be addressed by using a sample of parameter sets instead of a single parameter set (e.g., Vrugt et al. 2003). Beven and Binley (1992) proposed the Generalized likelihood uncertainty estimation (GLUE) framework to address the uncertainty in model identification by looking for models that are equally acceptable (often referred to as equifinal).

The total uncertainty in streamflow forecasts can be assessed using hydrological ensembles that sample from all uncertainty sources listed above. For example, Refsgaard and Storm (1996), Krzysztofowicz (2001) and Kavetski et al. (2006) presented methods to sample the total uncertainty in hydrological modeling chains. Such approaches might be used to identify the relative role of each uncertainty source. Thiboult (2016) found that accounting for the uncertainty in initial hydrological conditions contributed to accuracy and dispersion for shorter lead times in an es, whereas for longer lead times, meteorological forcing uncertainty was important to maintain reliability. Demargne et al. (2010) showed that the hydrological model uncertainty, i.e., model parameters, model structure, and hydrological initial condition, were more significant for shorter than longer lead times. Especially for

high streamflow events, the meteorological uncertainty was most important beyond two days ahead. Also, Zappa et al. (2011) indicated that meteorological forcing was the dominant source of uncertainty in the hydrological forecasting chain.

In this thesis, the focus was on the uncertainty caused by meteorological forcing by using meteorological ensemble forecasts to establish hydrological ensemble forecasts. Therefore, hydrological forecast uncertainty and techniques applied to improve streamflow and flood forecasts were assessed, all within the Norwegian flood forecasting chain.

## 2.2  Meteorological ensemble forecasts

Dynamical weather systems are complex, non-linear and eventually chaotic, and hence it is not possible to fully describe and resolve their physical and dynamical evolution. Assumptions and simplifications are made to parameterize processes, while numerical approximations are made for discretization of continuous equations (Vannitsem et al. 2018). In addition to uncertainty originating in model description and parameterization, weather predictions are subject to the uncertainty in the initial conditions, data assimilation, and the chaotic nature of the atmosphere. This was acknowledged, for example, by Leith (1974) who suggested using Monte Carlo techniques to provide a range of alternative forecasts for the same future. Today, ensemble prediction systems are used to represent the uncertainties in weather forecasts. Singular vectors (Toth and Kalnay 1997) or breeding vectors (Molteni et al. 1996) are used to account for the errors in the initial conditions, whereas methods like the stochastically perturbed parameterization tendency (SPPT) scheme (Buizza et al. 1999; Palmer et al. 2009; Leutbecher et al. 2017) account for the uncertainty in the model's physical description and parameterization. Figure 5 shows the concept of initial condition perturbation and the SPPT used to establish the European Center for Medium-Range Weather Forecast (ECMWF) ensemble forecasts (ENS).
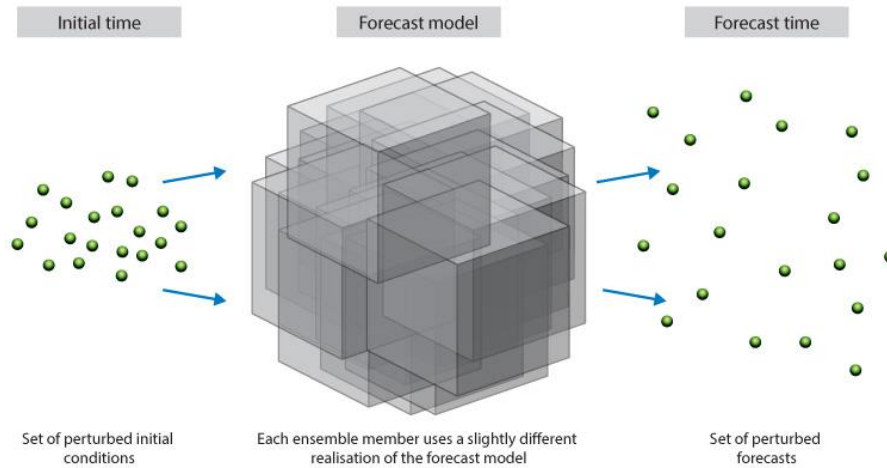
**Figure 5**. *The process of perturbing initial conditions and the stochastically perturbed parameterization tendency (SPPT; ECMWF 2017).*

Due to computer demand and the consequent high cost, there is a tradeoff between ensemble size and resolution when running global weather forecasting models. Studies have shown that shorter lead times gain the most by higher resolution, whereas larger ensemble size is beneficial to longer lead times (e.g., Miller et al. 2010; Buizza 2010; Kay et al. 2013).

Ensemble forecasts have two main advantages (Kalnay 2003). Firstly, the average of the ensemble is more skillful than any single member for forecasts beyond a few days. Secondly, reliability can be attained from the ensemble forecast. Reliability describes how well the ensemble is able to capture the observations, e.g., for 90% of the time, the observations should be within the 90% interval of the ensemble forecasts. A reliable forecast with a narrow (sharp) distribution is more informative than one with a broad distribution that makes it more difficult to separate events from nonevents (e.g., Gneiting et al. 2007; Leutbecher et al. 2017).

However, the ability of weather forecasting models to predict future states (predictability) is limited (Lorenz 1963, 1969). Predictability depends on atmospheric conditions and is higher for stable high-pressure situations, e.g., in the center of an anticyclone, compared to unstable atmosphere, e.g., the development of a frontal wave (Persson 2015). There are also regional differences, and predictability is generally higher for the extratropics than for the tropics (e.g., Haiden et al. 2018). In addition, due to data assimilation schemes which

are part of the dynamical weather and climate models, areas with high observation densities are less uncertain than areas with low observation densities (ECMWF 2020a).

## 2.3 Pre- and postprocessing

The grid resolution of the meteorological global ensemble forecasts is generally too coarse to be used directly as input for hydrological models at the local scale, which can affect the hydrological response. Compared to local observations, the raw (unprocessed) ensemble forecasts are often biased and underdispersed. Whereas bias describes the difference between the ensemble mean and the observations, dispersion describes the spread of the ensemble members and is a property of the ensemble alone. Underdispersion is most evident at shorter lead times, and can be explained by the growth rate of the perturbations applied in the ensemble prediction system, which are optimized for the medium range and tend to be smaller compared to those of the "real" atmosphere for the short range forecasts (eg., Buizza et al. 1999). Moreover, the model resolution affects the details in the description of topography, which is important to how well e.g. orographic enhancement of precipitation and convective precipitation is resolved. Along a coastline, a detailed representation of the coastline and hence the separation between land and ocean cells are important, and will affect e.g. radiation schemes and temperature estimates. Therefore, raw ensembles are often seen as unreliable in a statistical sense (Buizza 1997; Wilson et al. 2007).

For the reasons explained above, meteorological forecasts are downscaled from the grid resolution provided by the weather forecasting model to a scale appropriate for hydrological modelling. Downscaling includes dynamical and statistical approaches; dynamical downscaling usually involves a regional weather model providing forecasts at a high resolution able to resolve small-scale processes. A non-hydrostatic convective-scale weather prediction model like Applications of Research to Operations at Mesoscale - Meteorological Cooperation on Operational Numerical Weather Prediction (AROME-MetCoOp; Müller et al. 2017) provides an explicit representation of relevant physical descriptions, for example by improving orographic representation and allowing for convective processes. However, dynamical downscaling is computationally expensive and

not readily available; statistical methods are, therefore, an easier and more practical choice. Statistical downscaling defines a statistical model that fits the ensemble forecasts to the observation and provides adjusted forecasts that are statistically calibrated (e.g., Hamill and Colucci 1997; Persson 2015).

A selection of statistical ensemble processing methods is presented, for example, in Li et al. (2017) and Vannitsem et al. (2018). The methods differ in their sensitivity to length of training data and ensemble size and how spread and bias are corrected. To achieve improved streamflow forecasts, processing can be applied to the input (i.e., precipitation and temperature), output (i.e., streamflow), or all variables. Table 1 shows a selection of studies in which pre- or postprocessing was used for the end purpose of improving streamflow forecasts.

Ensemble model output statistics (EMOS; Gneiting et al. 2005; Wilks and Hamill 2007; Vannitsem et al. 2018) are a group of regression methods in which the conditional mean of the distribution is a linear combination of the ensemble members and a nonconstant variance is allowed. The most applied variant of the method uses a normal distribution (non-homogenous Gaussian regression; Wilks and Hamill 2007). Another popular approach is Bayesian model averaging (BMA; e.g., Raftery et al. 2005; Wilson et al. 2007), which combines the weighted predictive distribution of each model (or ensemble member). The weights can be considered the prior probability attached to each model (or ensemble member). BMA is widely used to improve the spread–skill of bias-adjusted (calibrated) ensemble forecasts for a range of variables, and is applied using different probability distributions (kernels).

Raftery et al. (2005) applied BMA to temperature and sea-level pressure using a normal density function. A Bernoulli–gamma distribution is suggested for precipitation (Sloughter et al. 2007). The gamma distribution was found less appropriate for streamflow (Vrugt and Robinson 2007) and wind (Sloughter et al. 2010). Friederichs and Thorarinsdottir (2012) used generalized extreme value distribution to process peak wind. Duan et al. (2007) applied the Box–Cox transformation to streamflow to achieve a normal distribution in a multimodel BMA approach. A number of studies have shown the applicability of BMA

(see Table 1 for examples) in which the choice of an appropriate kernel distribution, e.g. normal for temperature and a Bernoulli-gamma for precipitation, is important for the results.

Application of BMA to precipitation generally increases the performance of the forecast compared to that of unprocessed forecasts, but less so for high-precipitation events (Sloughter et al. 2007; Liu and Xie 2014). The skill of streamflow forecasts increases when BMA is applied to a multimodel streamflow dataset (Duan et al. 2007; Parrish et al. 2012). Xu et al. (2019) further showed that by applying BMA to streamflow forecasts, it is possible to improve the forecasts to levels similar to those of streamflow forecasts that account for several sources of uncertainty (e.g., forcing, initial conditions, and hydrological models; Thiboult et al. 2016). The flexibility of BMA is appealing, and is relevant to studies in which combinations of hydrological models and different sources of forcing can be evaluated.

**Table 1.** *A selection of studies that show different pre- and postprocessing approaches for the purpose of improving daily ensemble streamflow forecasts. The table provides an overview of the processing approach used in the studies (Approach), which variables that has been applied a processing technique (precipitation (P), temperature (T), streamflow (Q)), short description of catchments studied (Catchments.), whether or not floods have been evaluated (Flood), and a short summary of the results (Findings). The table is sorted from studies that include pre- and postprocessing to those that only applies postprocessing.*

| Reference | Title | Catchments | LT | Perfor. metric | Approach | P | T | Q | Flood | Findings |
|---|---|---|---|---|---|---|---|---|---|---|
| Zalachori et al. 2012 | Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. | Catch: 11 Area from 220 to 3600 km². No information on snow or regimes | 1-7 | RMSE, BS, RPS, BSS, RPSS, PIT-histogram | Analog Past error | X | X | X | N | Results showed that correction of meteorological forcing was important. However, accounting for hydrological uncertainties improves the quality of operational streamflow forecasts. Analog method used previous 50 similar events based on an evaluation of the 700 and 1100 hPa fields. Ensmeble dressing to remove bias and improve reliability |
| Benninga et al. 2017 | Performance of ensemble streamflow forecasts under varied hydrometeorologic conditions. | Catch: 1 Mountainous (957 km², mean streamflow 9.4 m3/s), snow, snowmelt in spring | 1-10 | CRPS/CRPSS/rank-histogram, sharpness, reliability, ROC/AUC | QM | X | X | X | Low- (2.7m³ /s), mediu m-, high-flow (10.4 m³/s) | The best performance was obtained for all lead times without any processing. Postprocessing reduced performance compared to preprocessing alone. QM: Quantile mapping |

| Reference | Title | Catchments | LT | Perfor. metric | Approach | P | T | Q | Flood | Findings |
|---|---|---|---|---|---|---|---|---|---|---|
| Roulin and Vannitsem 2015 | Postprocessing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions, and model errors. | Catch: 1 Area 317 km2, elevation 280–650 m, and mean annual rainfall 1029 mm | 1- 9 days | ME. RMSE, Spread, CRPS | ELR EVMOS INFL | X | | X X | N | Postprocessing of the hydrological ensembles improved the verification scores better than preprocessed precipitation ensembles. By combining the postprocessing of both precipitation and streamflow further improvements were achieved in cases of large precipitation biases. Extending logistic regression (ELR), error in variable model output statistics (EVMOS); variance inflation (INFL). |
| Sharma et al. 2018 | Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. | Catch: 4 (nested) From 380 to 12 362 km². Flooding in the cool season combined snowmelt and rainfall Summer convective events | 1–7 days | CRPS, CRPSS, BSS (flood threshold) | HCLR ARX QR | X | | X X | Y | Both pre- and postprocessing applied separately improved the streamflow forecast skill. The effect depended on the season. The combination of both ensured the highest skill improvements, but postprocessing alone was a sound alternative. Approaches: heteroscedastic censored logistic regression (HCLR), autoregressive model with a single exogenous variable (ARX), and quantile regression (QR). |

| Reference | Title | Catchments | LT | Perfor. metric | Approach | P | T | Q | Flood | Findings |
|---|---|---|---|---|---|---|---|---|---|---|
| Verkade et al. 2013 | Postprocessing ECMWF precipitation and temperature ensemble reforecasts for operational hydrological forecasting at various spatial scales. | Catch: 49 Rhine, 48 sub basins Snowmelt swiss alps impact Rhine in spring, | 1 – 10 days | BS, CRPS, ROC (AUC), RME | QT LR LogR | X X | X X | | N | Unconditional bias in forcing was resolved by a simple quantile-to-quantile transform (QT). There were moderate improvements in conditional bias and skill of the forcing ensembles, and the improvements varied with forecast lead time, amount, and spatial scale. However, there was no direct translation of improvements in forcing data into streamflow forecast skill. Linear regression with an assumption of bivariate normality (LR) and logistic regression (LogR). |
| Bogner et al. 2016 | Postprocessing of streamflows in Switzerland with the focus on low flows and floods. | Catch: 2 Sihl (hourly data floods, few events) Thyr (daily data lowflows) | 1-5 (Cosmo-LEPS) | QuantileScore, CRPS, Q-Q plot MAE, NS | AR WT QRNN | | | X X X | Y | For floods, the QRNN method with log-normally distributed quantiles performed best in terms of validation period. However, no clear results were obtained for the hindcasts applied to test the models in a forecast situation. AR was the less good approach for all metrics and levels. Few events and no clear improvements Autoregressive (AR), wavelet transformations (WT, Bogner 2011), quantile regression neural networks (QRNN). |
| Fundel and Zappa 2011 | Hydrological ensemble forecasting in mesoscale catchments: sensitivity to initial conditions and value of reforecasts. | Catch: 3 47, 186, and 1696 km². 2 alpine with snow | 1-10 days. ECMWF IFS | Spread, RMSE, BSS, CRPSS, Economic value. | ExtLR | | | X | 80th quantile | Forecasts calibrated with ExtLR reduced forecast bias for most lead times and catchments. Forecasts benefited from calibration on reforecasts (including alpine catchment and snow). Extended logistic regression (ExtLR; Wilks 2009) avoids possible inconsistencies in the calibrated probabilistic forecasts for different exceedance thresholds by ensuring decreasing forecast probabilities with increasing threshold. |

19

| Reference | Title | Catchments | LT | Perfor. metric | Approach | P | T | Q | Flood | Findings |
|---|---|---|---|---|---|---|---|---|---|---|
| Reggiani et al. 2009 | Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. | Catch: 1 Rhine, 160.000 km$^2$, HBV, ECMWF ENS | 1-9 days | RPSS, spearman rank correlation, economic value | Bayesian processor | | | X | N | The postprocessor verification showed good performance when the ranked probability skill score was compared to non-Bayesian uncertainty assessment, like ranking threshold exceedance probabilities for the members of a streamflow ensemble prediction. The Bayesian processor delivers a posterior probability density function of the expected flow rates or water levels, conditional on all information available from the outset of the forecast. |
| Xu et al. 2019 | Hydrological postprocessing of streamflow forecasts issued from multimodel ensemble prediction systems. | Catch: 38 Canada, high snowmelt flow April, autumn preciptaion | 1-9 days | KGE, SSP, MCRPS | BMA-multimodel | | | X | N | BMA was used to postprocess multimodel hydrological forecasts. The overall quality of the forecasts improved by maintaining ensemble dispersion with lead time. Skill improved even though the schemes only included two sources of uncertainty. |

## 2.4  Observed climate change

Flood frequency and magnitude are sensitive to changes in precipitation and temperature. In Europe, the spatial patterns of floods are changing both in terms of timing and magnitude (Blöschl et al., 2017, 2019). An analysis of daily extreme precipitation in the second half of the 20th century showed an increasing trend for most of Europe; however, there were regional and seasonal differences and the highest increase in extreme precipitation is detected in stations with an overall increase in precipitation (e.g., Klein Tank and Können 2003). During the 20th century, the average temperature increase in Norway was similar to the global increase of 0.7 °C (Hanssen-Bauer and Nordli 1998), and precipitation in the same period increased by 5–15% (Hanssen-Bauer and Førland 1998). The highest precipitation increases averaged over the whole of Norway is about 18%, seen from the 1970s. The increase is largest for autumn in southern and southeastern Norway and for spring in western and northern regions (Hanssen-Bauer et al. 2017). The observed changes in temperature and precipitation have an impact on floods. Vormoor et al. (2016) showed that there is a decreasing trend in floods in northern Norway due to reduced snowmelt, whereas the trend is increasing in southern and western Norway, caused mainly by an increase in intense rainfall events.

## 2.5  Global climate projections and regional downscaling

To assess changes in a future climate, studies depend on emission scenarios representing anthropogenic external forcing. Meehl et al. (2009) described the difference between daily forecasts and climate projections; daily forecasts are basically an initial condition problem, whereas multi-decadal to century climate change projections are a forced boundary condition problem. When the external forcing is known, global climate models (GCM) validated with historical data estimate rather well the forced response at a global scale (e.g., IPCC 2010).

The variability and uncertainty of climate change at the local scale are larger than at the global scale (e.g., Hawkins and Sutton 2009). To assess the local or regional hydrological

21

impact, it is necessary to translate global model output to local values. The discrepancy in resolution is often solved by dynamical downscaling (regional climate models [RCM]) and statistical downscaling. Independent of downscaling method, it is considered essential to establish the local data used for impact studies. Statistical methods rely on calibration of the model that fits the GCM output to observations in an historical reference period, and it is often assumed that the statistical relationships between the global and local data stand in a future changing climate (e.g., Dixon et al 2016). For dynamical downscaling, the computer demand for running high-resolution models is high, and international initiatives have been established to provide climate projection data from multiple models and resolutions available for scientific use, e.g., the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor et al. 2012) or the European Coordinated Regional Downscaling Experiment (EURO-CORDEX; Jacob et al. 2014; Hazeleger et al. 2015). The horizontal model resolution of the RCMs in the EURO-CORDEX is typically about 12 km (0.1°), whereas the horizontal resolution of the driving GCMs is 100 km (1.25°) or more.

In a future climate, precipitation intensity is projected to increase globally with a more pronounced increase in mid and high latitudes (e.g., Meehl et al. 2007). For Norway, the projections show that the increase in precipitation observed in the last century would continue in the future (Hanssen-Bauer et al. 2017). The annual increase in the median value for precipitation was 3–14% at the end of the century under the RCP4.5 emission scenario and based on the CMIP5 12-km mulitmodel ensemble. However, there was a large spread in the model simulations, with the northern regions having the highest uncertainty (spread). Especially important for floods is that the number of days with high precipitation and high intensity of precipitation would increase, and consequently, the possibility of rain-induced floods. Assessment of floods in a future climate indicates that both the number and magnitude would increase in the western regions of Norway, whereas snowmelt-induced floods in the eastern and high-elevation areas would decrease (e.g., Beldring et al. 2006; Roald et al. 2006; Sorteberg et al. 2018; Hanssen-Bauer et al. 2017; Vormoor et al. 2015; Lawrence and Hisdal 2011).

For a plausible assessment of floods in a future climate, it is important to represent high precipitation events correctly. It is generally acknowledged that a high-resolution model is

necessary for simulation of extreme precipitation events. Moreover, the topography must be well represented for accurate orographic simulation. However, the computer costs of high-resolution modeling are often a limiting factor. A storyline approach is presented below that allows for high-resolution modeling, and an alternative approach is introduced to assess potentially high-impact extreme flood events in a future climate.

## 2.6 Storylines

Multi-model GCM modeling approaches, in which probabilities are attached to different scenarios, is the traditional choice for estimating future climate change. However, the IPCC (2010) pointed out instances where different approaches are preferable. For example, the best model for the purpose should be chosen for cases in which specific weather processes are better described and resolved by one GCM compared to others. High-resolution modeling at a high computational cost is required to resolve certain events. In such cases, storylines following selected physically plausible trajectories (events) can be an alternative to the coarse-resolution GCM ensemble (Hazeleger et al. 2015). Storyline approaches are currently being advanced to complement information obtained from traditional probabilistic assessments. A storyline can be tailored to unfold past or plausible future events in a physically self-consistent way. A storyline based on plausibility rather than probability can be used to raise risk awareness of specific events and thereby enable decision-makers to adapt proactively (Shepherd et al. 2018; Shepherd 2019). There are obvious limitations to the storyline approach; since limited datasets are used, care must be taken in selecting events, and the validity of models and the purpose of the study must be transparent (Sillmann et al. 2019).

Including a full and comprehensive modeling chain is important in the storyline approach; for flood impact studies, at least one hydrological model should be included. Each catchment has its specific properties, and a linear relationship does not necessarily exist between precipitation and streamflow. Schaller et al. (2016) showed that even if there is a signal in GCMs and RCMs, it can be dampened in the hydrological modeling. Therefore, conclusions on hydrological impacts should not be drawn from extreme precipitation events alone (Felder et al. 2017).

An example in which a storyline is applicable is the landfall of atmospheric rivers in a future climate. For such cases, it is essential to have a model resolution that can describe the topographic barrier (Neiman et al. 2009). Moreover, the model should be able to resolve the relevant processes, e.g., the small-scale sharp frontal structures that are associated with the development of atmospheric rivers (Whan et al. 2020). Zappa et al. (2013) demonstrated that CMIP5 models with a higher horizontal resolution were more effective at representing both cyclone number and intensity and storm-track position and orientation. Dominguez et al. (2018) applied a storyline approach to assess how a devastating atmospheric river event in 2007 might appear in a warmer future climate. They used a "pseudo-global-warming" approach and applied the climate change data to a modelling chain to reveal the atmospheric-hydrologic-hydraulics and economic impacts of the "future" event. However, no studies have provided a high-resolution simulation of future plausible atmospheric river events and followed the extreme precipitation event all the way to floods.

## 2.7 Hydrological forecasting versus climate projections

There are obvious differences in forecasting streamflow and assessing climate change impact of flood events. Forecasting is an initial condition problem and one important key to the success of weather and flood forecasts are observations used in assimilation schemes to update model states (ECMWF 2020a). Climate change is more a boundary condition problem, and emission scenarios and process parameterization become even more critical to represent the future climate (Meehl et al. 2009). Forecasts can be verified/evaluated due to the relatively short time horizon, whereas the climate change projections cannot be verified in a similar way. We can assess how well the models performs for current climate and thereby assume they have a similar performance for the future climate.

In both weather forecasting and climate projection studies ensembles are used to enable alternative realizations for a given time horizon and displays uncertainty from different sources in the modelling chain. In weather forecasting uncertainties in initial conditions are accounted for by initial condition perturbations. In climate projections different emission scenarios are used to capture the uncertainty in future environment, whereas

deficiencies in a model structure or process descriptions can be represented by multi-model ensembles or alternative parametrization schemes. Both the initial condition perturbation and a stochastically perturbed parameterization tendency (SPPT, Leutbecher et al. 2017) is used to establish the ensemble of weather forecasts at ECMWF, and thereby assessing a probability of future weather. The same SPPT approach can be applied within a storyline approach to represent deficiencies in model parametrization, and thereby establish an ensemble of plausible weather events at the end of the century, without any probabilistic interpretation. Independent of the approaches used to create an ensemble, the ultimate goal is to improve the weather forecasts or the climate projections, and to better capture the uncertainty attached to prediction in both the near and the far future.

The resolution of the model determines which processes that can be resolved and how well topography can be described. Due to the coarse resolution of global weather and climate models, the model output is often biased when compared to the local observations. These biases in model output are troublesome for hydrological modeling. The model outputs therefor need to be downscaled and bias corrected, and similar methods are used for forecasting and climate projections. Regional climate models are used to get a better representation of regional weather and climate. To provide forecasts and projections that are calibrated and representative of the statistics of observations, different statistical approaches are applied to account for systematic and random errors in predictions. Well adopted correction techniques should be applied for improved streamflow and flood simulations at the local catchment scale.

## 2.8 Gap of knowledge

In flood forecasting, ensemble forecasts add value since they provide a probability attached to the exceedance of flood thresholds. Available processing techniques to remove bias and improve dispersion, and thereby improve forecasting skill, are described in e.g. Vannitsem et al. (2018). To improve streamflow forecasts most studies have focused on precipitation and streamflow. However, for Norway and other snow prone regions, temperature is important to estimate snow accumulation and snow melt, both important for correct streamflow forecasts. There are few studies that have looked at improvements in

forecasting skill for the unique and combined effect of preprocessing temperature, precipitation, and postprocessing streamflow for forecasting both floods and all streamflow values (table 1). There are no conclusive results to which variables processing should be applied, nor to preferred technique. Examples of findings are that streamflow forecasting skill will depend on flow levels (e.g., Benninga et al. 2017), lead time (e.g., Demargne et al. 2010; Thiboult et al. 2016), and for different seasons (e.g., Monhart et al. 2019). Several studies underline that to provide robust results, there is a need to test processing techniques for more catchments, more data, and events. Only two of the studies listed in table 1 use more than 30 catchments (Verkade et al. 2013; Xu et al. 2019) where Verkade et al. (2013) focused on preprocessing only and Xu et al. (2019) focused on postprocessing. None of these two studies analyzed the effects on floods. It is therefore a need for more knowledge on the combined effect of preprocessing temperature and precipitation and postprocessing streamflow on flood forecasts for a large sample of catchments. In this thesis, these knowledge gaps are addressed by defining the research questions Q1 and Q2 defined in the Introduction (Section 1.2) and answered in paper I and paper II. A flood forecasting chain consisting of 145 Norwegian catchments was used. The catchments represent different regions, varies in size, climatology, and hydrological regimes, and the study period includes several flood events affecting most parts of Norway. In the processing setup, different techniques are evaluated, for the single variable and for combinations. A special focus is laid on how the applied techniques affect floods, and spatial and temporal patters in predictability across lead times were investigated.

Floods are expected to change in a future climate. For Norway, studies indicate an increase in rain induced floods, whereas the large snowmelt driven floods are expected to decrease at the end of the century (e.g., Hanssen-Bauer et al. 2017). Atmospheric rivers are one of the dominant flood-generating weather systems onto western Norway and are responsible for the most damaging flood events in the region (Stohl et al. 2008). Even though the future change in atmospheric rivers have gained attention in climate studies (e.g., Ralph and Dettinger 2011; Espinoza et al. 2018), the ultimate effect on floods at the catchment level is not well studied. In this thesis research questions Q3 and Q4 are addressed in paper III and paper IV. An event-based storyline approach (Hazleger et al 2015) is used to provide

a high-resolution modeling alternative to assess the impact of future extreme precipitation events. One argument for applying a storyline is to raise the awareness of risk by focus on plausibility rather than probability (Shepherd et al 2018). One problem with assessment of atmospheric river induced floods is that the extreme precipitation is enhanced by orographic lifting, and therefore a high horizontal model resolution is essential for a good representation of topography. Most global climate models run at a coarse resolution not well suited to describe orographic precipitation and nor the atmospheric river induced floods. In this thesis the entire modelling chain comprises a high-resolution global climate model, a regional weather prediction model, and a catchment based hydrological model. The weather and hydrological models are in operational use, and therefor optimally calibrated to the region of interest, here 37 different catchments over western Norway are included. This enables an insight into the flood impact of future atmospheric river events for a larger region. Further, the use of known flood reference, for example flood exceedance thresholds and warning colors are similar to those in operational use, mitigates any mismatch between the perception of climate change between the climate scientist and the user of the data.

# Chapter 3
# Study area and data

## 3.1 Norwegian climatology and catchments

The Norwegian topography and diversity of climatic zones (Hanssen-Bauer and Förland 1998; Hanssen-Bauer and Nordli 1998) create large variations in precipitation and temperature patterns, resulting in a variety of hydrological regimes (Gottschalk et al. 1979). Steep mountains combined with a temperate oceanic climate cause high annual precipitation in western coastal Norway (Fig. 6, left). Inland eastern and northern areas of Norway experience less precipitation and a larger difference between winter and summer temperatures compared to the coastal areas (Fig. 6).

Hanssen-Bauer et al. (2017) gave an overview of the past, current, and future climate in Norway. In the assessment of climate change in Norway, a multimodel ensemble for each of the emission scenarios RCP4.5 and RCP8.5 was included. The RCP4.5 projections indicate a temperature increase of about 2–4 °C with the highest increase in the northern regions, especially in winter (Fig. 7, right). RCP8.5 (Fig. 8) indicates an overall higher temperature compared to RCP4.5, while the relative temperature increase is highest in the winter/cold season. The percent change in precipitation is higher under RCP8.5 than under RCP4.5; however, the spatial precipitation patterns of the two scenarios are similar.
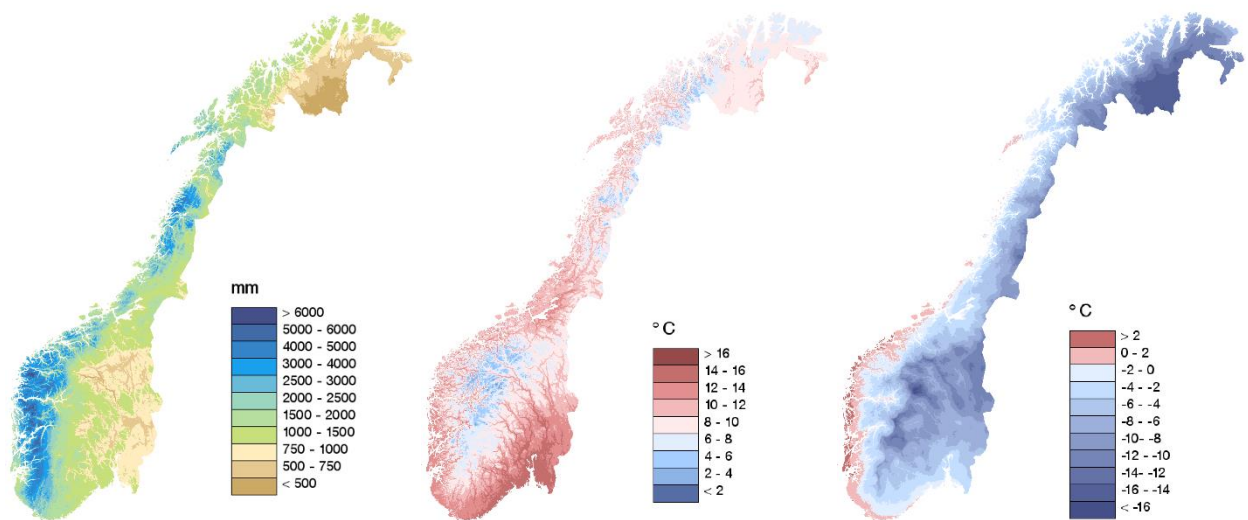
*Figure 6. Mean annual precipitation in mm (left), mean annual summer temperature in °C (middle), and mean annual winter temperature in °C (right). All data were for the reference period 1971–2000 (Norwegian Centre for Climate Services [NCCS],2020).*
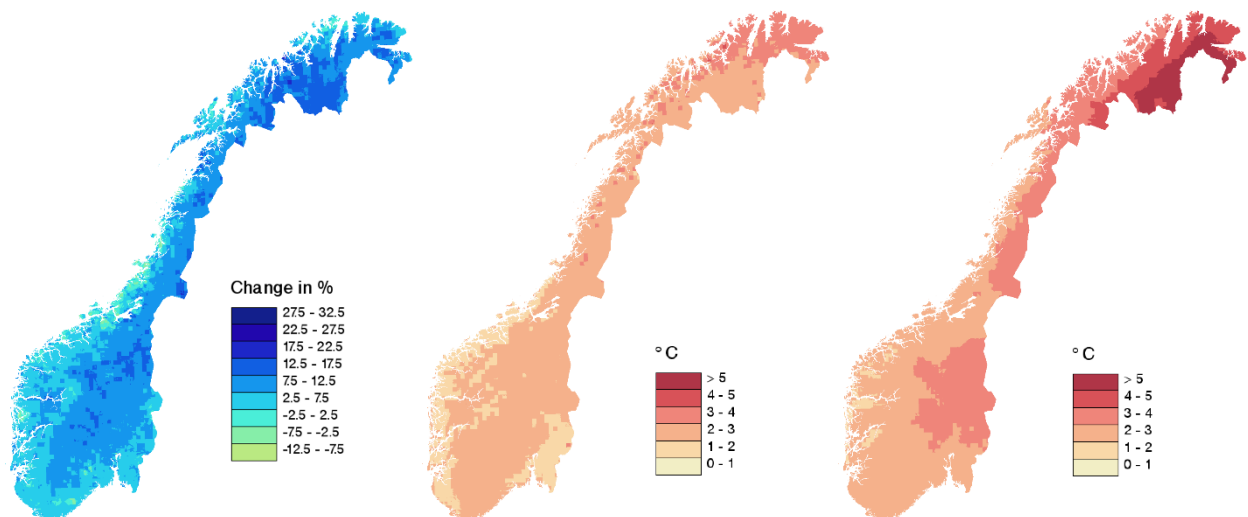


*Figure 7. Changes in precipitation (left) as percent change at the end of the century (2071-2100) compared to the reference period, change in mean summer (middle), and winter temperature (right) as degrees Celsius. The reference period was 1971–2000 and the simulations were under RCP4.5 (Norwegian Centre for Climate Services [NCCS] 2020).*
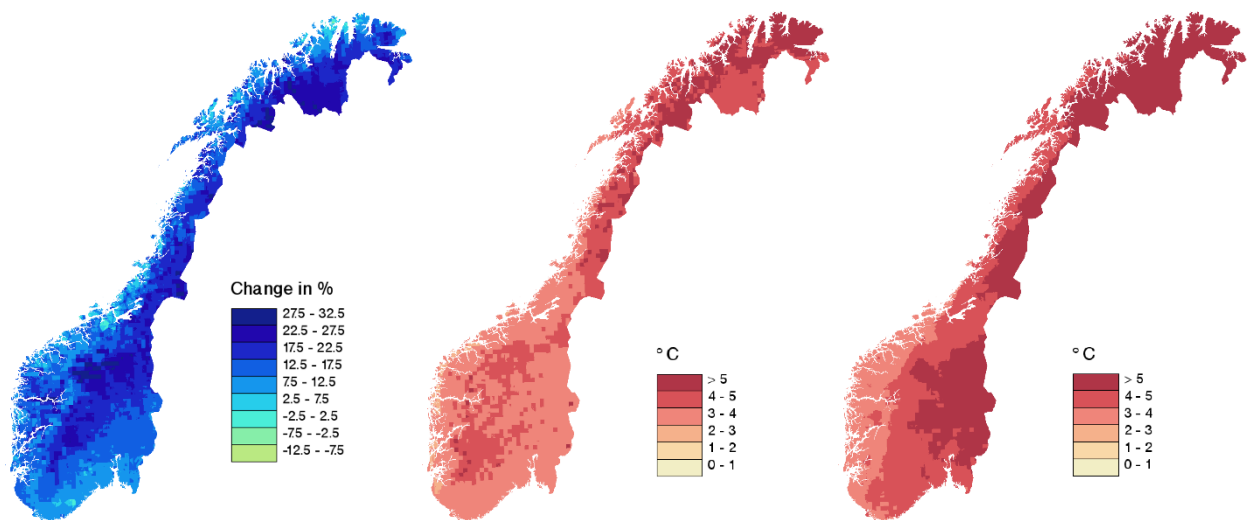
*Figure 8. Changes in precipitation (left) as percent change at the end of the century (2071-2100) compared to the reference period, change in mean summer (middle), and winter temperature (right) in degrees Celsius. The reference period was 1971–2000 and the simulations were under RCP8.5 (Norwegian Centre for Climate Services [NCCS] 2020).*



*Figure 9. Catchments included in the operational flood forecasting system, partitioned into regions; north (N), mid (M), west (W), south (S), and east (E).*

The flood forecasting system in Norway is based on streamflow forecasts for 145 catchments, representative of most regions and climates (Fig. 9). The catchments vary in size from 3 to 15447 km$^2$ and in elevation from 103 to 2284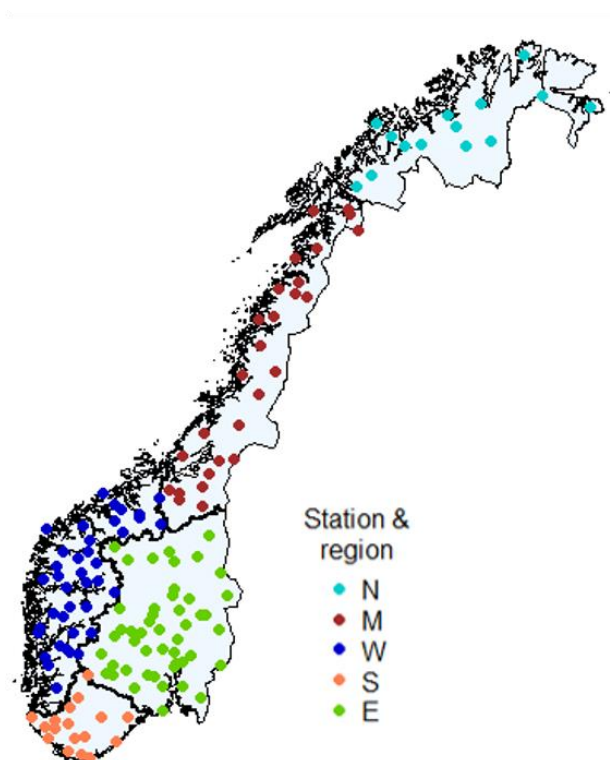 m. The catchments are mostly unregulated, and can be divided into five regions, north (N), south (S), west (W), mid (M), and east (E), which reflect major hydroclimatological zones (Hanssen-Bauer et al. 2017; Vormoor et al. 2016). Despite the regional and seasonal variations in temperature, most catchments under the present climate experience subzero temperatures and seasonal snow cover. Any snow stored during winter is responsible for the low flows in winter and the possibility of snowmelt-induced floods in spring. In Norway, two basic hydrological regimes can be defined. High flows during the autumn and winter due to heavy rainfall are typical of the coastal regions (S, W, and M), whereas the inland and north (E and N) regions more often have high streamflow during the spring due to snowmelt. However, many possible transitions exist between these two basic streamflow patterns within each region (e.g., Gottschalk et al. 1979).

## 3.2 Meteorological observation: SeNorge

In this study, the gridded SeNorge version 1.1 data (Tveito et al. 2005; Tveito 2007; Mohr 2008) were used as temperature and precipitation references in Papers I and II, and in modeling chain 2 (see Figure 3) to establish the initial model conditions, and as forcing for the HBV model to calculate the simulated reference streamflow in all papers. The SeNorge dataset is a 1×1-km interpolated grid based on the available station data at each time step and provides daily data back to 1957. The interpolated temperature is created by detrending the observed 24-h mean temperature to sea level. The temperature interpolation accounts for location, including latitude and longitude. The difference in elevation within a 20-km radius from the measuring point gives an indirect measure of topography, and affects vertical temperature changes (Tveito 2002; Tveito et al. 2005; Mohr 2008). The gridded SeNorge precipitation data use precipitation corrected for undercatch. The interpolation is based on a triangulation approach that takes into account the station elevation and a digital elevation map of Norway (see Mohr [2008] for details). Since most measuring stations are

placed at low elevation close to populated areas, the observational points for higher elevations and sparsely populated areas are fewer and are hence more uncertain.

## 3.3 Meteorological forecasting: ECMWF ENS

This study used the ECMWF temperature and precipitation forecast ensemble (ENS) for the period March 3, 2013 to December 31, 2015, with an original grid resolution of 0.25° or approximately 32 km (i.e., model cycles/versions 38r1/2, 40r1, and 41r1; ECMWF 2020). The ENS consists of a control forecast for which the forecast model is run using the best initial conditions, model description, and model parameters. The 50 additional members of the ENS are generated by adding small perturbations to the initial conditions and to the model physics schemes, subsequently running the model with different perturbed conditions compared to the control forecast. The ensemble represents the weather forecast uncertainty, with ideally a small spread in the ensemble indicating a stable atmosphere and less uncertain forecasts compared to an unstable atmosphere resulting in a large spread. A more detailed description of the ECMWF ENS system is provided, for example, in Buizza et al. (1999, 2005) and Persson (2015). The ECMWF ENS was used in modeling chain 1 in Figure 3 to generate retrospective streamflow forecasts.

### 3.3.1 ECMWF IFS cycle - to choose a representative model version

The ECMWF forecasting models consist of a set of applications that are available in the Integrated Forecasting System (IFS; ECMWF [2020c]) and undergo changes relatively often. In 2016, there was a major upgrade that changed the horizontal and vertical resolution as well as introducing changes to the radiation schemes (IFS 41r2). The higher resolution and the radiative scheme solved to a large degree some of the temperature biases that were apparent along the Norwegian coast (Seierstad et al. 2016). Moreover, an earlier experimental version of the IFS 41r1, which included a higher horizontal resolution, improved the precipitation intensity data of the extreme events that affected western Norway in October 2014 (Figure 2 in EFAS 2014).

Based on the findings above, it was anticipated that the new IFS cycle 41r2 would improve the weather forecasts, and ultimately the hydrological simulations. An experimental dataset

was used in which IFS 41r2 was rerun for two extreme precipitation events affecting Norway in 2014 (original ENS, IFS 40r1) and 2015 (original ENS, IFS 41r1), thereby enabling a comparison of the same events of the original IFS ensemble and the new ensemble run by IFS 41r2. Both chosen events caused extensive flooding in affected catchments, and the floods were not well captured using the original ensemble. The first event in 2014 improved the forecasts using IFS 41r2 for 50% of the affected catchments. IFS 41r2 for the second event, which affected southern Norway in 2015, improved the forecasts for all catchments. However, flood size was underestimated for all catchments, and neither the original nor the IFS 41r2 ensemble forecasts were able to predict the observed floods, see for example Stordalsvatn in Figure 10.
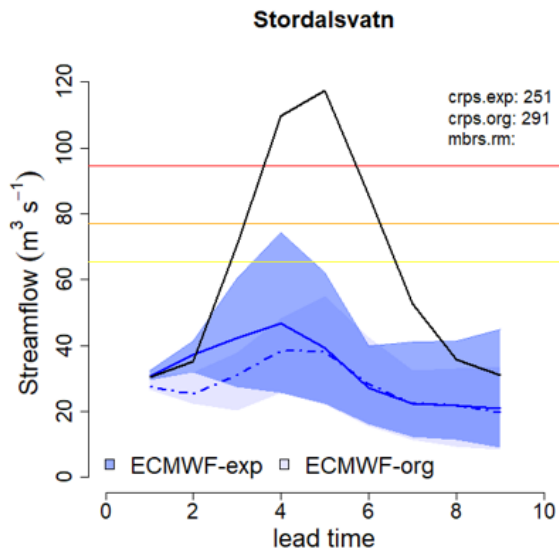


*Figure 10*. An example of the performance of the two IFS versions for the 2014 atmospheric river event in the Stordalsvatn catchment. The issue date was 5 days before the peak of the event. crps.exp = continuous ranked probability score (CRPS) for the experimental ensemble (IFS 41r2); crps.org = CRPS for the original ensemble (IFS 40r1). The black line is the reference flood (simulated using observations), blue line is the median of the experimental ensemble (the ensemble range indicated by blue transparent color), and the dotted blue line is the median of the original ensemble forecast (the light blue transparent color indicating the range of the original ensemble).

The comparison of IFS cycles showed that the new model cycle improved the streamflow forecasts. However, the improvement was not sufficient to provide reliable forecasts for the selected events; therefore, some processing seemed appropriate also for the new IFS cycles. Since this study focused on floods, the earlier dataset covering a period with many

floods was chosen instead of the resent dataset for a period consisting of unusually few floods.

## 3.4 Future climate projections

### 3.4.1 EC-Earth

The global climate model EC-Earth is based on the ECMWF integrated weather forecasting system IFS cycle 31r. EC-Earth version 2.3 (Hazeleger et al. 2010; Haarsma et al. 2013) provided the data used to evaluate the impact of future climate atmospheric rivers on western Norway (modeling chain 3 in Figure 3). EC-Earth is a high-resolution global model (resolution of about 25 km, T799L91). By comparing the integrated water vapor transport (IVT) of the model simulations to IVT of the ERA-Interim (Dee et al. 2011), Whan et al. (2020) showed that the model is well adapted to represent both the intensity and frequency of atmospheric rivers reaching the west coast of Norway. The RCP4.5 emission scenario (van Vuuren et al. 2011) was used to represent a future climate. A detailed description of the model setup is provided by Haarsma et al. (2013), and the model has been validated and used in several studies (e.g., Haarsma et al. 2013; Bintanja et al. 2014; Baatsen et al. 2015; Van Haren et al. 2015; Van der Linden et al. 2018). In this study, two extreme precipitation events were selected from a present-day scenario between 2002 and 2006 and a future scenario between 2094 and 2098. A perturbation method similar to the operational method used at ECMWF (SPPT, Figure 5) was applied to the selected events. By rerunning EC-Earth for all perturbed events, an ensemble of 10 members were created for each original event.

### 3.4.2 AROME-MetCoOp

The non-hydrostatic weather forecasting system AROME-MetCoOp (Müller et al. 2017) is the operational short-range (1–3 days) forecasting system for Norway, Sweden, and Finland. In the operational setup, the AROME-MetCoOp model was initialized and forced at the lateral boundaries by ECMWF IFS, which was replaced by EC-Earth for the setup used in Papers III and IV. The high-resolution (2.5 km) temperature and precipitation forecasts from AROME-MetCoOp are used as input for the operational flood forecasting

systems in Norway. Although the ECMWF IFS is able to represent reasonably well the large-scale precipitation patterns for extreme events, precipitation intensities and spatial distribution are better represented by AROME-MetCoOp; this is mainly explained by the better representation of orographic precipitation by the model's spatial resolution (Müller et al. 2017).

In this thesis, the operational weather prediction model AROME-MetCoOp was used in modeling chain 3 (see Figure 3) to downscale the extreme events selected from the high-resolution EC-Earth GCM. In the operational setting, AROME-MetCoOp has been shown to give a better spatial representation of extreme precipitation than ECMWF (Müller et al. 2017) and is, therefore, suitable for the purpose of assessing flood impact of atmospheric river events in a future climate.

# Chapter 4

# Methods

## 4.1 Modeling framework

A simplified presentation of the models and the pre- and postprocessing techniques used in the different studies of this thesis are shown in Figure 11. In the operational flood warning setup, both ECMWF-HRes and AROME-MetCoOp were used as forcing, but for different lead times. The warning levels were defined by the exceedance of thresholds estimated from the reference streamflow.
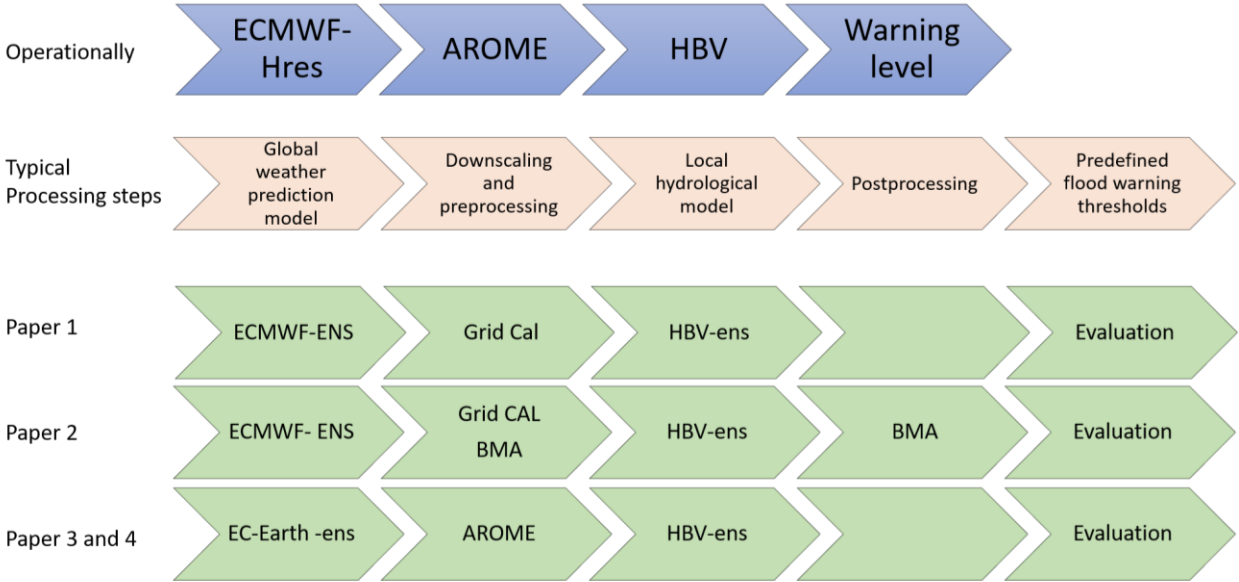


***Figure 11****. Simplified operational flood forecasting chain: Top row shows the data and models used in the Norwegian operational chain, followed by typical models and processing steps, and last the three approaches applied in Paper I-IV.*

In Papers I and II, different statistical pre- and postprocessing techniques were implemented and evaluated for all the data and for floods. For the future flood impact storylines (Papers III and IV), the regional weather prediction model AROME-MetCoOp was used to downscale and improve the spatial representation of weather for extreme events, and thereby potentially improve the flood simulations.

## 4.2 Precipitation–runoff model

The Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström 1976; Sælthun 1996; Beldring 2008) is a conceptual precipitation–runoff model implemented in the operational Norwegian flood-forecasting service. The HBV catchment scale, which uses daily average temperature and precipitation, was selected to estimate streamflow in this thesis. In the HBV model, each catchment is subdivided into ten equally large elevation zones. Each zone accounts for vertical precipitation and temperature gradients. Snow, soil moisture, and groundwater processes were described by separate modules and represented the water balance in the catchment by

$$P - E - Q = \frac{d}{dt}[S_{SP} + S_{SM} + S_{GW} + S_{SW}],$$

where P was precipitation, E was evapotranspiration, Q was streamflow, $S_{SP}$ was snowpack storage, $S_{SM}$ was soil moisture storage, $S_{GW}$ was groundwater storage, and $S_{SW}$ was soil water storage.

For calibration and validation of the hydrological model, SeNorge version 1.1 precipitation and temperature data described in Section 3.2 is used as forcing, whereas the streamflow observations are from the hydrological database of the Norwegian Water Resources and Energy Directorate (NVE; https://www.nve.no/hydrology/). The Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) and volume bias are used as calibration and validation metrics for the operational models (Ruan Gusong, personal communication, June 15, 2016). The Nash–Sutcliffe efficiency (positively oriented with an optimal value of 1) averaged for all catchments for the calibration period (1980 to 1995) is 0.74 with zero volume bias, and 0.71 with 2.2% volume bias for the validation period (1996 to 2012). This thesis applied the same setup and parameters as those of the operational flood-forecasting model.

To avoid biases and uncertainties originating from the hydrological model, it is common practice to evaluate the effect of the processing scheme on streamflow by using a reference forecast, which is a simulated streamflow from the hydrological model forced with meteorological observations. This approach was adopted in this thesis, and SeNorge v1.1 was used to establish the reference streamflow (see Figure 2(2), reference streamflow Qo).

## 4.3 Pre- and postprocessing

The meteorological and hydrological ensemble forecasts were pre- and postprocessed by including statistical processing approaches, which are described in more detail below.

The raw forecasts were downscaled by applying a nearest neighbor interpolation. Nearest neighbor was chosen primarily since it is less computer demanding compared to, for example, bilinear interpolation. Whenever no other preprocessing techniques were applied, temperature was adjusted using a standard atmospheric lap rate of 0.65 °C/100 m to account for the discrepancy between the different grid elevations (Sheridan et al. 2010; Persson 2015).

### 4.3.1 Grid calibration - MetNorway

This thesis applied the operational preprocessing methods used by the Norwegian Meteorological Institute (MET Norway); all methods are available at https://github.com/metno/gridpp/. The calibration parameters used at the time of the study period were provided by MET Norway. The processing schemes were applied to the raw ensemble forecasts and for the study period 2013–2015, and thereby established a new set of ensemble forecasts referred to as the calibrated ensembles ($T_{cal}$ and $P_{cal}$ in Figure 2). The following is a summary of the methods used by Met Norway to establish the calibration coefficients.

MET Norway established the calibration parameters for temperature by applying quantile mapping (Seierstad et al. 2016; Bremnes 2007) using both ensemble ENS re-forecast (Owens and Hewson 2018) and Hirlam (High Resolution Limited Area Model; Bengtsson

39

et al. 2017) gridded temperature forecast. Mapping is applied to the gridded forecasts, and for each grid cell. Monthly unique quantile transformation coefficients are determined by using data from a three-month window centered on the target month, e.g., the May analysis consists of April, May, and June. The same coefficients, based on mapping the first 24 hours, are applied to all lead times and members. A 1:1 extrapolation is used for forecasts outside the observation range.

The method applied to precipitation is mainly targeted at improving the spread of the ensemble, and 200 WMO station observations are used to establish the parameters. For precipitation, a Bernoulli–gamma distribution is applied to account for the non-continuous nature of daily precipitation. A logistic regression is used to describe the discrete part of precipitation (i.e., rain or not), and uses the ensemble mean and the fraction of members with less than 0.5 mm precipitation as predictors. The continuous part is described by the gamma distribution with the mean and variance as predictors. To fit the bias model, the cube-transformed mean precipitation is used, whereas the untransformed mean forecast is used to fit the variance model.

In this thesis, calibrated temperature ensemble forecasts were obtained for all issue dates by using the calibration parameters that MET Norway applied operationally at the time (approx. 2013 to 2015). The data were, therefore, similar to what would have been available for operational weather forecasting at that time. For the CAL methods, the corrections were applied to each ensemble member. Keeping the order of the members ensured consistency between the calibrated temperature and precipitation members, which is important for hydrological modeling. By applying the available meteorological methods, it was possible to evaluate whether the improved skill in the weather forecasts propagated to the streamflow and flood forecasts.

### 4.3.2 Bayesian model averaging

BMA is an adaptive method for pre- and postprocessing, and is used to correct spread errors in bias-corrected ensemble forecasts (e.g., Wilson et al. 2007). This thesis adopted the BMA approach suggested by Fraley et al. (2010), which includes a simple bias adjustment applied to each ensemble member using a linear regression. The output from the BMA is

a probability density function (pdf), which is a weighted average of the pdfs centered on each of the corrected forecast members (eq. 1). For each ensemble member, the pdf (referred to as the kernel) for the quantity to be forecast, $y$, is denoted by $f(y|x_m)$. For an ensemble consisting of $M$ members, $x_m$ is the forecast by ensemble member $m$, and the density function conditioned on all $M$ ensemble members is the weighted average of the pdf for each member:

$$f(y|x_1, \dots, x_M) \sim \sum_{m=1}^{M} w_m f(y|x_m) \qquad 1.$$

where $\sum_{m=1}^{M} w_m = 1$ and the weights are interpreted as the posterior probabilities of each ensemble member. When the ensemble members are indistinguishable (exchangeable), as is the case with ECMWF ENS, all members are given equal weights, and the BMA mean is treated as a constant for all members (Fraley et al. 2010). To account for the specific properties of temperature, precipitation, and streamflow, different distributions were used as kernels (see below for details).

Different training lengths were evaluated to train the BMA models. A training length of 45 days, meaning the 45 daily forecasts and observations prior to the forecast issue date, was used to establish the BMA parameters. Unique parameters were hence obtained for each issue date, catchment, and lead time.

### 4.3.3 Probability density distributions

The normal distribution (eq. 2) was used as the kernel for preprocessing temperature (Raftery et al. 2005) and postprocessing Box–Cox transformed streamflow (Box and Cox 1964; Duan et al. 2007). The mean was specified as $a_0 + a_1 x_m$, where $x_m$ was the temperature forecast for ensemble member $m$ and $a_0$ and $a_1$ were regression parameters to account for any bias. The variance $\sigma^2$ was defined as the same for all ensemble members. Then, the pdf for the forecasted variable $y_t$ conditioned on all ensemble members was

$$f(y_t|x_1, \dots, x_M) \sim \sum_{m=1}^{M} w_m N(a_0 + a_1 x_m, \sigma^2) \qquad 2.$$

For postprocessing streamflow, the BMA with a normal kernel, similar to eq. 2, was applied, where forecasts $x_m$ was replaced by the Box–Cox transformed streamflow $q_m^*$.

A Bernoulli–gamma distribution was used to model precipitation as suggested by Sloughter et al. (2007). The pdf for the forecasted variable, $y_p$, was established for each member in eq. 3 and combined the probability for zero precipitation ($f(y_p = 0|x_m)$) and the pdf for the amount of precipitation ($h(y_p|x_m)$). The second part of eq. 3 described the pdf of precipitation, given that it was larger than zero. The $I_{\{*\}}$ was unity if the condition in brackets was true and zero otherwise,

$$f(y_p|x_m) = f(y_p = 0|x_m)I_{\{y_p=0\}} + f(y_p > 0|x_m)h(y_p|x_m)I_{\{y_p>0\}} \qquad 3.$$

For the indicator function for ensemble member $m$, the probability of zero precipitation was given by eq.4,

$$f(y_p = 0|x_m) = \frac{1}{1 + exp\left(b_0 + b_1 x_m^{1/3} + b_2 \delta_m\right)} \qquad 4.$$

where $b_0$, $b_1$ and $b_2$ were regression parameters common to all ensemble members, and $\delta_m$ equaled unity if $x_m = 0$ and zero otherwise. Following Sloughter et al. (2007), the cube-root transformation of precipitation ($y_p$) was assumed to follow a gamma distribution in which the mean ($\mu_m$) and variance ($\sigma_m^2$) of the distribution depended on the original forecast ensemble member ($x_m$) as described in eq. 5.

$$\mu_m = c_0 + c_1 x_m^{1/3} \text{ and } \sigma_m^2 = d_0 + d_1 x_m \qquad 5.$$

The parameters $c_0$ and $c_1$ $d_0$ and $d_1$ were the same for all ensemble members. In the case of temperature, the BMA-pdf for precipitation was specified as the weighted average of the pdf for each ensemble member.

### 4.3.4 Rank correlation maintaining temporal and intervariable consistency

In addition to preprocessing the meteorological variables to achieve downscaled and calibrated forecasts applicable for use in hydrological models, it is important to preserve physical and temporal consistency between the variables. In this thesis, the consistency between members and lead times were preserved in the gridded calibration approach. $T_{cal}$ and $P_{cal}$ were produced by applying the processing schemes to each ensemble member separately. However, the BMA models define a probability density function from which 51 new ensemble members are randomly drawn for each lead time and variable independently. The temporal trajectory and the consistency between temperature and precipitation is thereby lost. An empirical copula approach (Wu et al. 2018) can be used to reconstruct a structural intervariable relationship for multivariate postprocessing. There are different strategies to establish a reference structure. The Schaake shuffle (e.g., Clark et al. 2004) uses observations to create the reference structure, whereas ensemble copula coupling (ECC, Schefzik et al. 2013) uses the structure of the raw ensemble. This thesis applied a simplified version of the copula approach. It applied a transfer function defined by the ranked correlation between the original ensemble members and the new BMA-processed values. The catchment average temperature and precipitation values for all ensemble members were used to establish the transfer functions. For each variable and lead time, the ensemble member with the highest value from the original ensemble was reassigned the highest value from the new BMA-processed ensemble, the original member with the second highest value was reassigned the second highest value, and so on. Consequently, it was possible to keep temporal and intervariable consistency for each of variable.

## 4.4 Evaluation of performance

The performance and skill of the applied processing schemes are assessed by appropriate evaluation measures. Different measures will enhance specific characteristics of the forecasts, and the choice of performance measure may affect the choice of model, model setup, parameter values, or processing schemes (Jolliffe and Stephenson 2012).

### 4.4.1 Reliability

Rank histograms (e.g., Hamill and Colucci 1997) are useful visual aids to establish the reliability of ensemble forecasts. The observations are ranked with respect to the corresponding ensemble forecast, and the shape of the rank histogram can provide insights on the strengths and deficiencies of the ensemble forecast system. A U shape indicates underdispersion in which a larger than expected part of the observations falls in the tails of the distribution, whereas a convex shape indicates overdispersion. Sloped rank histograms indicate systematic forecast bias (Hamill, 2001). Decomposition of the chi-squared test statistics for the rank histogram is suggested by Jolliffe and Primo (2008), and provides numerical values to describe the slope (bias) and convexity (dispersion) of the rank histogram. This approach was used in Paper I.

Paper II assessed reliability by a Q–Q plot in which the cumulative relative frequencies of observations from the rank histograms were plotted against the cumulative uniform distribution. These plots enabled comparison of the reliability of all processing schemes for each catchment and evaluation of bias and under- and overdispersion. A perfect reliable forecast will follow the 1:1 line.

### 4.4.2   The continuous ranked probability score and skill score

The continuous ranked probability score (CRPS; Hersbach, 2000) measures the distance between the ensemble forecast and a reference observation. The ensemble forecast is expressed by the cumulative density function, and if the forecast is a single value, CRPS equals the mean absolute error. CRPS has the same dimensions as the evaluated variable, and is the preferred score to assess the performance of ensemble forecasts. CRPS measures the distance between the observation $x_o$ and the ensemble forecast, where the latter is expressed by the cumulative density function $F_x(x)$:

$$CRPS(F_x, x_o) = \int_{-\infty}^{\infty} [F_x(x) - \mathrm{H}(x - x_o)]^2 dx \qquad 6.$$

where H is the Heaviside function, which is zero when the argument is less than zero, and one otherwise.

44

Skill scores are convenient to compare the relative performance of variables or between catchments since these scores are dimensionless. The continuous ranked probability skill score (*CRPSS*; eq. 7) is calculated by including a benchmark (*CRPS_B*) that the new forecast should outperform.

$$CRPSS = 1 - \frac{CRPS}{CRPS_B}$$

7.

The daily climatology data were used as a benchmark in Paper I. A climatological temperature ensemble of 55 members were created from the daily catchment mean values from each day of the year, and similarly, daily streamflow climatology was established from the reference streamflow calculated by the HBV model, forced with 55 years of temperature and precipitation. The focus of Paper II was evaluation of different processing schemes, and the raw streamflow forecasts were used as the benchmark. This allowed evaluation of the relative improvement of streamflow forecasts by applying different processing schemes.

### 4.4.3   Contingency tables, hits, and false alarms

The occurrence and nonoccurrence of floods are binary events that can be summarized in a contingency table (table 2).

**Table 2.** *The contigency table gives an overview of hits (H), missed events (M), false alarms (F), and correct nonevents (N), and is used to evaluate the performance of a forecasting system.*

|  |  | Occurrence of event | |
|---|---|---|---|
|  |  | No | Yes |
| Forecasted event | No | *N* | *M* |
|  | Yes | *F* | *H* |

For an ensemble system, a hit is a certain probability or a defined number of ensemble members exceeding a certain threshold. Since floods are rare events, the number of events

is small compared to the number of nonevents. If correct negative warnings are included as successful forecasts, the evaluation would result in a high score with little information about the performance during floods. A high hit rate is the main goal; however, it should not be at the cost of issuing too many false alarms. The critical success index (CSI; Donaldson et al. 1975; Jolliffe and Stephenson 2012) is a score that accounts for and will, therefore, penalize the hits by both the missed events and the false alarms (eq. 8),

$$CSI \;\; = \frac{H}{H + F + M} \qquad\qquad 8.$$

Therefore, the CSI was chosen as the method to evaluate the performance of the processing schemes in Paper II to specifically mimic the operational value of processing schemes for multiple catchments during flood conditions.

## 4.5 Evaluation of future floods

In previous studies, the probabilities of floods, the change in seasonality, flood generating processes, and the changes in return levels have been assessed by multi-model ensembles including different GCM, RCM, bias correction methods, and emission scenarios (e.g., Beldring et al. 2006; Lawrence and Hisdal 2011; Hanssen-Bauer et al. 2017). This thesis focused on a modeling chain that is specifically well adapted to assess the impact of future floods caused by ARs. High-resolution modeling is computer demanding, and for this reason, the most extreme events were selected. The selection represented plausible events with the highest precipitation within the study area and were further perturbed to assess some of the model uncertainty. Model output was downscaled to the resolution of the impact modeling and consistency between the variables was kept.

The event-based storyline approach was used in Papers III and IV to infer changes in flood impact between present and future extreme events. The number of realizations that exceeded the different flood warning thresholds was compared between the present and future climate events. The 50-year flood (Q50) estimated for the present climate was used as a threshold. The largest floods from the observational record (60 years) were compared to the largest floods from the projection to see whether this level was exceeded at any

catchments by any realizations. Furthermore, the contribution to events of initial hydrological conditions relative to perturbations of the GCM was evaluated by calculating the relative mean absolute deviation caused by the ensemble and by the initial hydrological conditions.

# Chapter 5
# Results

This section provides a summary of the results of the four papers constituting this thesis. All studies were focused on the catchments of the Norwegian flood forecasting system. A total of 145 catchments were part of this study, although not all catchments were included in all sub-studies. The two first papers included catchments from all Norway and a variety of climatic regimes and events, whereas the last two papers focused on the impact of atmospheric rivers and selected catchments in western Norway. Most catchments were included in Paper I. The evaluation in Paper II included the catchments that exceeded the flood level in the two-year period of the analysis. In Paper III, the catchments within the western region, most susceptible to ARs, were investigated. In Paper IV, two catchments were highlighted to show the hydrological variability at different temporal and spatial model resolutions for a present climate and a future climate atmospheric river event.

## 5.1 Paper I: Streamflow forecast sensitivity to air temperature forecast calibration for 139 Norwegian catchments

Precipitation is the most important flood generating mechanism in most catchments. Snowmelt, which is controlled to a large degree by temperature, is another important flood generating mechanism for regions with seasonal snow cover. Most Norwegian catchments experience subzero temperature with seasonal snow cover during winter, and are hence prone to snowmelt floods in spring. The main objective of Paper I was to investigate the effect of temperature forecast calibration on the streamflow ensemble forecast skill.

Furthermore, the paper aimed to detect any regional patterns or catchment characteristics that influence the temperature and streamflow forecast skill.

This paper applied the temperature calibration approach used by the operational weather forecasting service to remove biases in the uncalibrated raw temperature forecasts ($T_{ens}$) to obtain calibrated forecasts ($T_{cal}$). By running the hydrological model with observed precipitation in combination with the uncalibrated ($T_{ens}$) and the calibrated ($T_{cal}$) temperature forecasts, it was possible to assess the sensitivity of streamflow to temperature calibration alone.

The results showed that temperature ensemble calibration affected both temperature and streamflow forecast skill for most catchments, but season and region affected the skill differently. Figure 12 shows the skill change between calibrated and raw temperature and streamflow forecasts for spring and autumn. In autumn, the skill was improved for both temperature and streamflow, which was not the case for spring in which some catchments had negative skill change values for both temperature and streamflow.
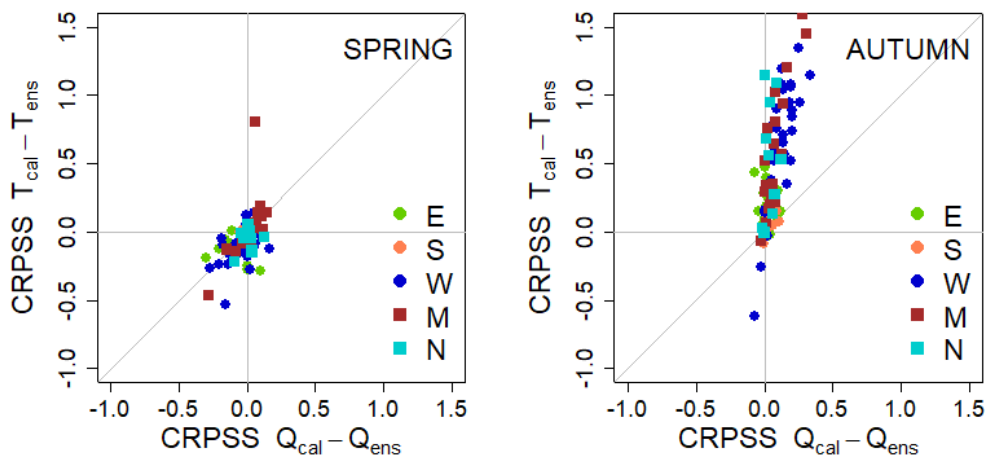


***Figure 12.*** *(Figure 5 in Paper I). Skill change for temperature (vertical axis) and streamflow (horizontal axis) for spring and autumn. Colors represent different regions in Norway: E = east; S = south; W = west; M = mid; and N = north. Grey line indicates the one-to-one relationship.*

The high skill improvements for temperature in the autumn resulted in improved streamflow forecast skill; however, the streamflow skill improvements were not of the same magnitude as the temperature skill improvements. This can be explained by subzero

temperature improvements that have less effect on streamflow. The cold bias in temperature was most apparent for the coastal regions in autumn and winter. Therefore, the effects of temperature calibration was dependent on both region and season.

In spring, although the temperature skill change was smaller than in autumn, the streamflow skill changes were more sensitive to temperature skill changes, apparent from a relationship closer to the one-to-one line in Figure 12. For temperature forecasts, it was useful to apply a regional division to identify spatial patterns in forecast skill. For streamflow, the spatial patterns were not so obvious, and it was not possible to be conclusive on any relationship between catchment characteristics and streamflow forecast skill.

In summary, the effect of temperature on forecast skill could not be neglected, and the streamflow forecasts were sensitive to the quality of temperature forecasts in climates with seasonal snow cover.

## 5.2  Paper II: The benefit of pre- and postprocessing streamflow forecasts for 119 Norwegian catchments

The main objective of Paper II was to evaluate the benefit of preprocessing, postprocessing, or a combination of the two compared to the raw ensembles. The performance of different combinations of pre- and postprocessing techniques was evaluated for ensemble forecasts within the framework of the Norwegian flood warning system. The ECMWF ensemble forecasts for temperature and precipitation were used to force the operational hydrological HBV model. Daily retrospective streamflow forecasts for two years (2014 and 2015) were estimated for 119 Norwegian catchments. Two approaches were used to preprocess the temperature and precipitation forecasts ($T_{raw}$ and $P_{raw}$): (i) the preprocessing applied to the gridded forecasts was the same as that used by the operational weather forecasting service in Norway, and included a quantile mapping approach for temperature and a zero-adjusted gamma distribution for precipitation, providing $T_{cal}$ and $P_{cal}$; and (ii) BMA was applied to the catchment average of temperature and precipitation, providing $T_{bma}$ and $P_{bma}$. BMA was used for postprocessing catchment streamflow forecasts, providing $Q_{bma}$.

Streamflow forecasts were generated for 14 schemes consisting of different combinations of the raw, pre-, and postprocessing approaches for the lead time of 1–9 days for the two-year period.

The forecasts for all streamflow and flood events were evaluated. The lowest flood warning level in Norway, the mean annual flood (QM), was used to separate the flood dataset from all streamflow. Evaluating all streamflow data showed that postprocessing improved the forecasts up to a lead time of two days, while using BMA preprocessing temperature and precipitation the forecasts were improved for 50–90% of the catchments beyond the two-day lead time. For flood events, preprocessing of precipitation and temperature gave better CRPS to marginally more catchments compared to the other schemes, but there was no obvious pattern. Postprocessing without preprocessing seems to be the least good approach. For longer lead times, postprocessing leads to poorer performance in increasingly more catchments, compared to using raw forecasts (see Figure 13)

In an operational forecasting system, warnings are issued when forecasts exceed defined thresholds, and confidence in warnings depends on the hits to false alarms and missed events ratio. Analysis of the CSI showed that many of the forecasts seemed to perform equally well. Furthermore, there were large differences in the ability to issue correct warning levels between spring and autumn floods. There was almost no ability to predict autumn floods beyond two days, whereas spring floods had predictability up to nine days for many processing schemes and catchments.

The results underline differences in the predictability of floods depending on season and flood generating processes, i.e., snowmelt-affected spring floods versus rain-induced autumn floods. Moreover, the results indicate that the ensemble forecasts are less good at predicting correct autumn precipitation; therefore, more emphasis should be given to finding a better method to optimize autumn flood predictions. To summarize, the flood forecasts benefited from pre-/postprocessing; however, the optimal processing approaches depended on region, catchments, and season.
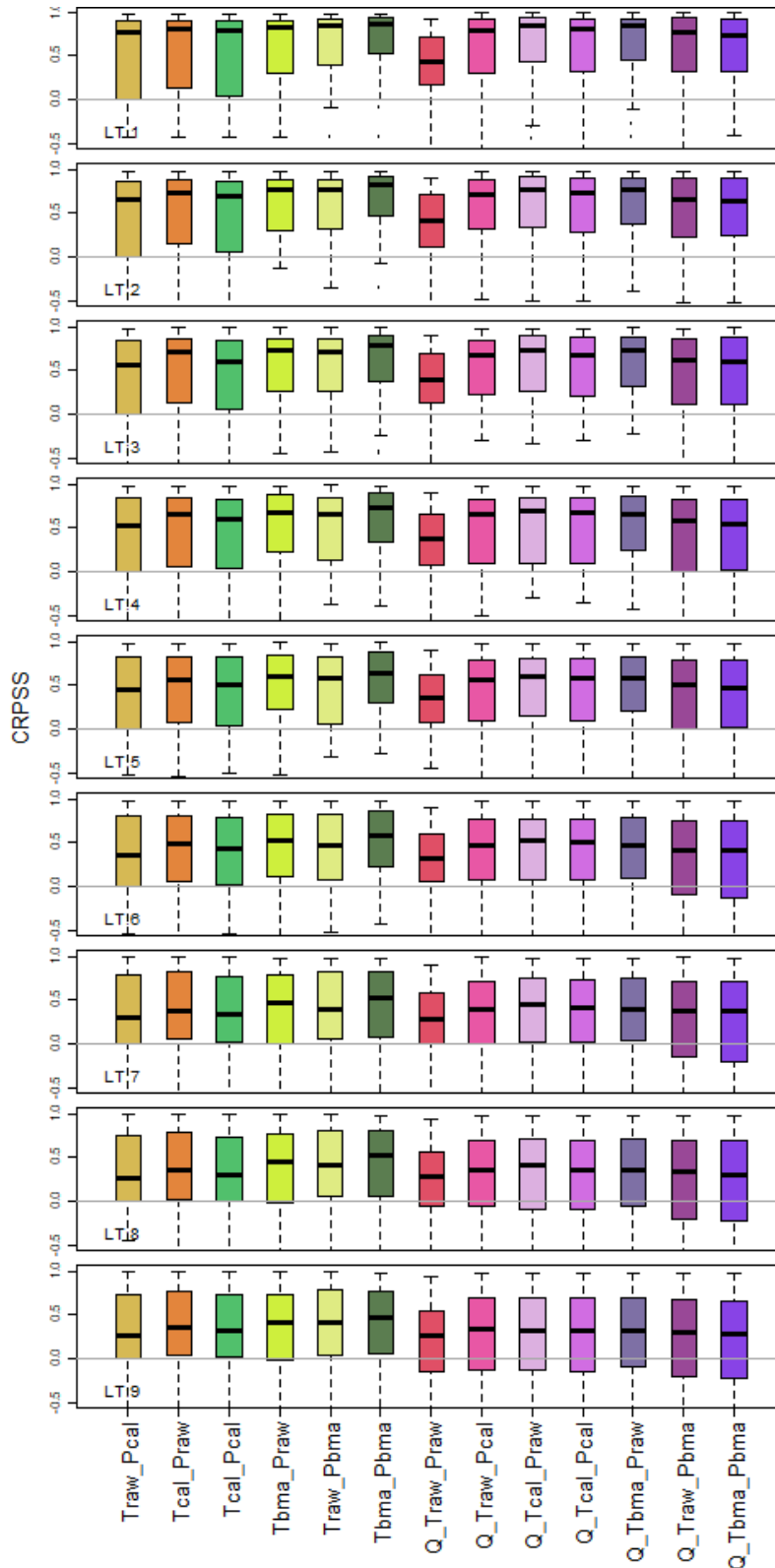
***Figure 13***. *(Figure 7 in Paper II). Boxplots of CRPSS for all catchments based on the flood dataset. All processing schemes (indicated on the horizontal axis) for all lead times LT:1 to 9 days (from top row and down). The first six boxplots show combination of preprocessed temperature (T) and precipitation (P), where raw=no processing, cal=grid calibration, and bma=Bayesian model averaging. The last seven boxplots are all postprocessed (indicated by the Q), with different combination of preprocessed temperature and precipitation..*

53

## 5.3 Paper III: An event-based approach to explore selected present and future atmospheric river-induced floods in western Norway

Future climate change at the global scale is well documented, whereas effects at the local scale are more challenging to model, and the estimated changes are more uncertain. Paper III investigated extreme precipitation events caused by atmospheric rivers and compared their flood impacts in a warmer future climate to those in the current climate by applying an event-based storyline approach. Furthermore, it evaluated the importance of initial hydrological conditions and meteorological forcing to the total flood impact.

Atmospheric rivers are specific events that are better resolved by high-resolution global models. Therefore, a modeling chain, which consisted of the high-resolution climate model EC-Earth and the regional weather prediction AROME-MetCoOp model, was created as a means for dynamical downscaling to provide better orographic representation of precipitation. To evaluate the flood impact, Paper III used the HBV model set up for 37 catchments located within areas that can be impacted by ARs.

Since the Norwegian catchments are at a high latitude and in steep terrain, parts of the catchment may be snow covered and experience snow fall controlled by elevation, which defines the spatial distribution of temperature. The role of the hydrological initial conditions were investigated using contrasting initial conditions from four different historical years that were chosen to represent a variety snow (snow and no snow) and soil moisture (dry and wet) conditions.

Thereafter, the magnitude and number of catchments affected by floods were analyzed. The main finding was that future events have a greater effect than present climate events on catchments with larger floods (Fig. 14). This finding is in line with expectations from previous climate projection studies of Norway. The new contribution of this study is that it visualized plausible impacts and showed how future events have a higher spatial impact, indicating that more catchments might be affected during one single event, which is an important finding for decision-making and flood prevention and adaption. Furthermore, perturbation of atmospheric model physics and parameters were of the highest importance

to assess a larger range of extreme precipitation events, whereas the initial hydrological conditions made an additional contribution to the total flood estimation.

Paper III showed that the storyline approach providing high-resolution modeling, similar to an operational modeling chain, is well adapted to describe atmospheric rivers in both the present and the future climate.
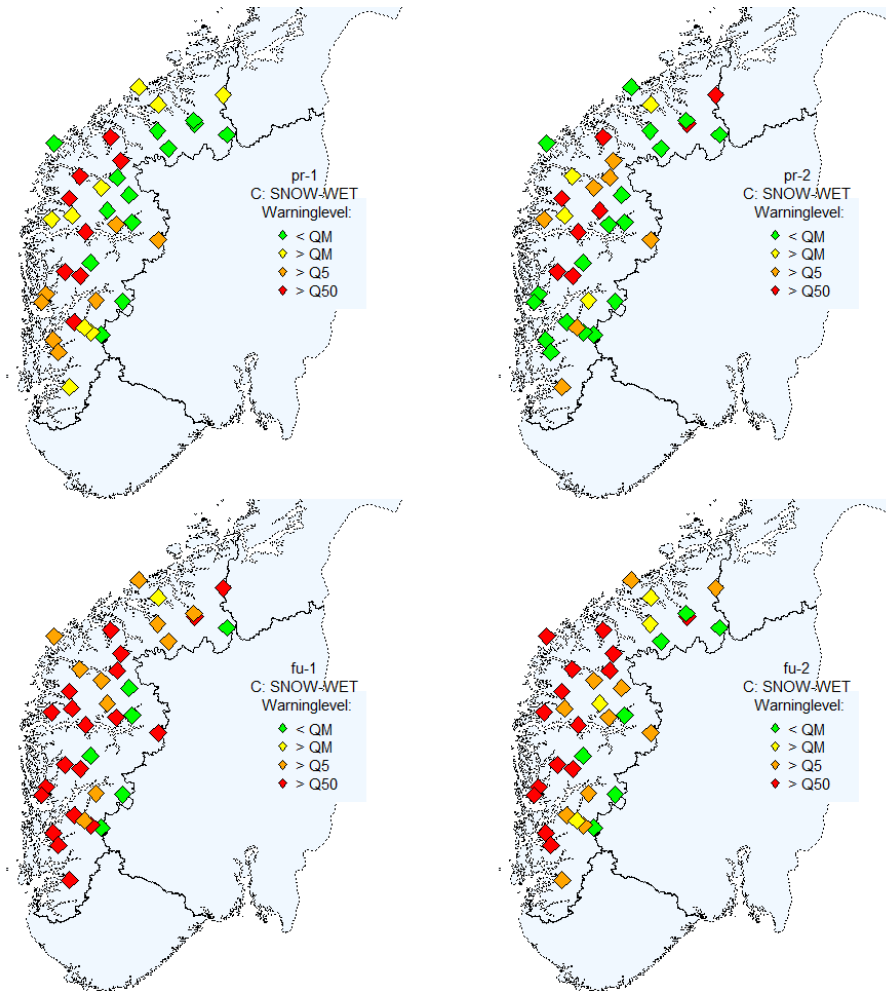


***Figure 14.*** *(Figure 9 in Paper III). Flood warning levels for catchments in western Norway for all four selected events. From top left, pr-1 and pr-2 (present climate) and fu-1 and fu-2 (future climate), using hydrological initial conditions that contains snow and high soil moisture (SNOW-WET). The highest warning level reached for each catchment is presented for all events. The yellow, orange, and red diamonds represent floods exceeding the warning thresholds of a median (> QM), 5-year (> Q5) or a 50-year (> Q50) return level, respectively. Green diamonds (< QM) indicate no floods.*

## 5.4 Paper IV: The role of spatial and temporal model resolution in a flood event storyline approach in western Norway

Paper IV (led by N. Schaller) explored the effect of high spatial and temporal model resolution when evaluating high-impact extreme events in a future climate and compared this to the impact in a present climate. The hypothesis was that a physical climate storyline approach, which allows for higher resolution modeling, can provide added value to existing probabilistic approaches. Paper IV applied the same dataset and was part of the same research project as the study presented in Paper III. Two catchments that had previously been severely affected by atmospheric rivers were chosen as the focus of the study, and the hydrological impact was investigated for different temporal and spatial model resolutions. The spatial resolution of the meteorological models was investigated by applying modeling chain 3 (Fig. 2), whereas the temporal resolution was investigated by HBV run by a co-author.

The topography of western Norway is challenging. Therefore, a high-resolution atmospheric model is necessary to improve the representation of orographically enhanced precipitation. In this study, precipitation intensity was realistically simulated with higher variability by the regional AROME-MetCoOp model compared to the global EC-Earth model. Furthermore, the higher resolution regional atmospheric model could model large precipitation differences in nearby valleys, whereas the coarser resolution global model provided a more uniform precipitation field.

The difference in precipitation between the regional and global model was reflected in the streamflow values (see Figure 15). Moreover, AROME-MetCoOp also affected the representation of temperature and will, therefore, have an additional effect on streamflow whenever parts of the catchments experience temperatures below zero. Furthermore, the hydrological model was able to better represent the culmination of flood events, and, therefore, the damage potential of the floods, when hourly rather than daily time resolution was used.
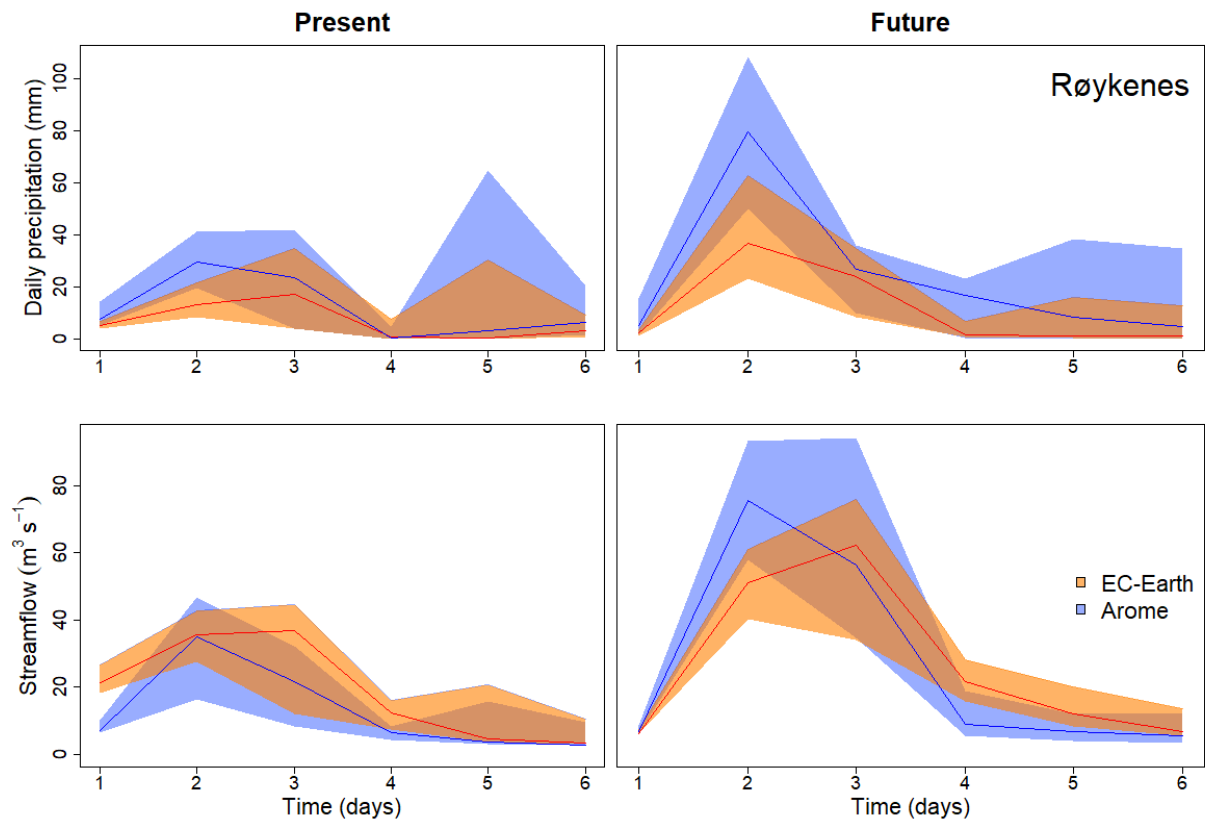
***Figure 15.*** *(Figure 3 in Paper IV). Present (left) and future (right) events in the Røykenes catchment. The top panels show daily precipitation and the bottom panels show streamflow. The shaded area represents the range of the 10 ensemble members for both.*

One drawback of this study was that only two catchments were evaluated. One of the catchments was not affected by the events, whereas in the second catchment, the future flood was clearly larger than the present one. For this study, a different selection approach, e.g., using the highest precipitation for the specific catchments, might have altered the events and, therefore, the impact on the catchments.

# Chapter 6
# General discussion and outlook

This thesis addressed two different topics: flood forecasting up to nine days ahead and assessment of floods caused by atmospheric rivers at the end of the century. Although these topics represent different challenges, there were similarities in the methodological frameworks used to address them. Similar modeling chains were used in which a global weather model was the first element. The resolution needed to force the hydrological impact models was essential to both forecasts and climate projections. Downscaling, either dynamical or statistical, was used to give a better representation of weather at the local scale. Statistical approaches (preprocessing) were used to downscale and calibrate the meteorological forecasts, whereas dynamical downscaling was used for the climate projections. Both approaches have common variations of techniques and models.

The time horizon is one of the profound differences between climate projections and weather forecasts. Climate projections do not need to maintain the accuracy expected of weather forecasts; however, the models used for climate projection should be validated and evaluated against historical observations to ensure that they are able to represent the climatology of a reference period. Furthermore, the projections should be plausible reflections of future trends and provide a statistically coherent representation of the future climate. On the other hand, weather forecasting should be able to provide reliable and accurate forecasts for a relatively short and limited time. Another difference is that weather forecasting is an initial condition problem for short lead times, where the chaotic nature of the atmosphere defines a limit of predictability (e.g., Lorenz 1963, 1969). For projections for the end of the century, the relative importance of different sources of uncertainty

depends on the scale. At the global scale, the emission scenarios are the main source of uncertainty, whereas at the local scale, model uncertainty is nearly as important as emission scenarios (Hawkins and Sutton 2009).

The following discussion first addresses the use of meteorological ensemble forecasts and different processing schemes for improving flood forecasts. Thereafter, storylines used to assess the impact of future flood events are discussed.

## 6.1  Ensemble flood forecasting

This thesis focused on the meteorological forcing uncertainty that makes the highest contribution to uncertainty, especially for flood forecasts (Zappa et al. 2011). The working hypothesis was that pre- and/or postprocessing would improve streamflow forecasts, as shown by previous studies (Table 1). A majority of these studies focused on postprocessing streamflow and one or two variables, and their processing only evaluated a few catchments. Only a few studies focused on floods. Papers I and II evaluated two years of retrospective forecasts for over 100 catchments with varying properties and climatic conditions. The catchments are representative of a high-latitude climate that includes seasonal snow cover. The number of catchments and the varied characteristics provide a unique insight into the robustness of approaches applied, especially how the applied techniques perform under flood conditions.

### 6.1.1 The full dataset - all streamflow

Paper I focused on assessing the effect of calibrated temperature forecasts on streamflow forecasts. The evaluation of raw temperature forecasts for Norway revealed a bias in the cold seasons, which was to a large degree removed by preprocessing. The temperature calibration affected streamflow both during autumn and spring, and typically related to the partitioning of snow and rain and to snow melt. Few studies have evaluated the ultimate effect on streamflow by processing temperature. Both Verkade et al. (2013) and Benninga et al. (2017) included temperature and precipitation in their preprocessing scheme, but they did not assess the effect solely of temperature on streamflow forecasts. A common outcome of these studies was that even though preprocessing precipitation and temperature

improved the skill of forcing, the effect was not translated into streamflow forecasts (Verkade et al. 2013; Benninga et al. 2017). A different result was obtained by Roulin and Vannitsem (2015) and Sharma et al. (2018) who found that preprocessing precipitation improved streamflow forecasts. By adding postprocessing to the evaluation, they were able to achieve further improvements. Zalachori (2012) reported that preprocessing alone did not improve the performance of streamflow for longer lead times, which was achieved by including postprocessing. However, postprocessing did not help to improve streamflow forecasts in Benninga et al. (2017). Most of the other studies agreed that postprocessing streamflow would improve streamflow forecasts for most lead times. All these contrasting results can be supported to some degree by the results presented in this thesis.

In Paper II, the processing schemes applied included two approaches for temperature and precipitation preprocessing, and one approach for postprocessing. The results of evaluation of the full dataset in Paper II showed that any processing scheme was better than the raw ensemble forecasts. These findings are in line with the expectation of most of the studies presented in Table 1, and the main hypothesis of this work. The results show that both temperature approaches, $T_{cal}$ and $T_{bma}$, perform well and provide improved streamflow forecasts when applied alone and in combination with postprocessing. For precipitation, $P_{bma}$ performs well, whereas for $P_{cal}$, both alone and in combination with $T_{cal}$, there are large differences in performance skill between catchments. This indicates that BMA applied to catchment values for precipitation is a better approach compared to the simpler grid calibration (CAL). Furthermore, postprocessing alone resulted in the lowest skill compared to preprocessing alone or the combination of the two. The improved performance by pre- and postprocessing is supported by other studies (Zalachory et al. 2012; Sharma et al. 2018; Roulin and Vannitsem 2015). Moreover, it is important to consider the performance of the original ensemble forecasts. For regions with low initial performance, the room for improvement is larger. This was shown by some large temperature improvements in regions with a large temperature bias. However, the relationship between improvement in temperature forcing and the resulting streamflow was not necessarily clear (e.g., Figure 12), mainly because streamflow is dependent on the combined effect of temperature and snow, or temperature and precipitation.

Another interesting result in this study is that for most of the catchments, postprocessing is less important as lead time increases, in contrast to the findings of Zalachori et al. (2012). This may be explained by the general size and memory of the catchments. In general, the Norwegian catchments are small, and a relatively quick response to precipitation can be expected. However, the catchments in Zalachori et al. (2012) are in the same size range as the Norwegian catchments; therefore, other aspects, for example, the techniques applied in addition to the initial performance of the raw ensembles, may be another explanation.

### 6.1.2 Floods

The difference in skill of the various processing schemes and catchments is less clear for the flood dataset compared to the full dataset. The BMA approach applied to precipitation revealed a weakness when preprocessing extreme precipitation events. For example, the preprocessed precipitation ensemble for the atmospheric river event in 2014 included some improbably high $P_{bma}$ values for some western catchments (e.g., Figure 12 in Paper II). Furthermore, in the examples above, $Q_{bma}$ suppressed the effect of the high precipitation values from $P_{bma}$. As a result, the combination of pre- and postprocessing ($P_{bma}$ and $Q_{bma}$) produced lower flood values compared to the other processing schemes. To avoid such incidents, techniques to constrain precipitation amounts or give a better representation of high precipitation values should be considered. For example, Ji et al. (2019) showed that by dividing the precipitation data into different precipitation categories, and thereafter training the BMA on relevant precipitation data, the forecasting skill of the probabilistic forecast was greatly improved. The variation in skill between the catchments for the flood dataset can be explained by the number of days included in flood evaluation, in addition to catchments characteristics, the nature of the events, and the season. Overall, the results showed that even though there were large differences between the catchments, only few catchments did not improve by any processing, and preprocessing $T_{bma}$ and $P_{bma}$ was marginally best.

Another interesting result is the distinct difference in predictability between spring and autumn floods, which to a large degree reflects the dominating flood-generating processes

in each season. Snowmelt dominates during spring, whereas rain dominates during autumn. Being more dependent on temperature, the snowmelt floods during spring are easier to forecast on longer lead times. The more intense autumn precipitation has a discontinuous spatial and temporal behavior, which may lead to large errors.

### 6.1.3 Outlook

For several of the flood events, an early signal of high precipitation is lacking, which is difficult to improve within this setup. However, there were events in which the storm path and the location of intense precipitation was misplaced. For such cases, it was not possible to correct the floods by using the processing techniques at hand. More flexible processing approaches that account for the uncertainty in spatial location of forecasted precipitation should be considered.

The global ensembles of medium range forecasts are designed to maximize the spread after 48 hours. A possible alternative to short lead times is the use of regional meteorological ensembles that outperform the global ensembles (Frogner et al. 2019a; Frogner et al. 2019b). However, such regional ensembles were not available for this study, but will be implemented in future studies. Other options are to combine the deterministic high-resolution ECMWF HRES forecast and the ENS within for example the BMA setup. The combination of the two systems provides more skillful precipitation forecasts than each system alone (Rodwell 2005/2006).

These findings show that for floods that are characterized by rapid changes in precipitation and streamflow, alternative approaches to choosing training data should be investigated, as well as including reforecasts and testing analog methods (e.g., Zalachori et al. 2012).

## 6.2 Future floods and storylines

The motivation for the studies presented in Papers III and IV was the atmospheric river flood events that caused extensive damage in the west coast of Norway in 2005 and 2014. Historically, atmospheric rivers have been responsible for the most damaging floods in this area (e.g., Stohl et al. 2008). Since the frequency and magnitude of atmospheric rivers are expected to increase in a future warmer climate (Whan et al. 2020), the main aim of the

studies was to investigate the flood impact of future atmospheric river events on the Norwegian west coast by applying an event-based storyline approach.

To ensure that chosen models are suited to the purpose of the study, model validation is especially important for the storyline approach. The high-resolution EC-Earth model has been used in several previous studies, e.g., to show how tropical hurricanes in a warmer future would be more prone to striking western Europe (Haarsma et al. 2013) and to assess changes in magnitude and frequency of atmospheric rivers (Whan et al. 2020). To assess changes in atmospheric rivers, previous studies were based on GCMs with a lower resolution compared to that of this study, and used specific signals of moisture and upslope winds to detect atmospheric rivers (e.g., Dettinger et al. 2011; Espinoza et al. 2018). Although these modeling approaches can determine changes in the intensity and frequency of atmospheric rivers, their resolution is not necessarily sufficient to represent the topography and hence the orographic lifting important for precipitation causing floods. The topography is well represented in the high-resolution EC-Earth model used in this thesis; Paper IV (Figures 1–3 in Paper IV) showed that there is good representation of precipitation and flood by the EC-Earth data. By implementing the regional weather forecasting model AROME-MetCoOp, the aim of this study was to further improve the representation of spatial intensity of precipitation. AROME-MetCoOp has been shown to improve operational precipitation forecasts compared to the coarser ECMWF global forecasts, especially for high precipitation large-scale events (Müller et al. 2017).

The results of Paper IV show that the cumulative precipitation of the larger region will not necessarily be increased by the AROME-MetCoOp model (Figure 1 in Paper IV), and for some events, it provided less precipitation than the EC-Earth model. This can be explained by the topographic representation that provides higher precipitation peaks due to the orographic effect (e.g., steep mountain), also provides areas in-between peaks (e.g., fiords) with less precipitation (Figure 2 in Paper IV). However, in hydrological modeling, the data from the AROME-MetCoOp model provides higher streamflow, indicating that the redistribution provides better representation of local precipitation within the catchments. An alternative to applying a regional model to downscale the global data is to apply statistical bias correction methods. Statistical methods are more feasible alternatives,

especially for large datasets when the computational cost is relevant. The statistical approach is highly dependent on the relationship between the model output (GCM or RCM) and the observations under present climate conditions. Therefore, it is important to define downscaling and bias correction methods that can account for and preserve the trends caused by climate change (e.g., Eum and Cannon 2017). By focusing on a few events in the storyline approach in this study, it was possible to run the AROME-MetCoOp model for selected events, which could not be done for the complete EC-Earth dataset. Consequently, this study provided physically consistent spatial precipitation and temperature fields at the highest available model resolution for Norway. In Paper III (Figure 5 in Paper III), the precipitation simulated by the AROME-MetCoOp model was compared to the observed precipitation for the 2005 atmospheric river event, and showed that the spatial pattern of precipitation is well represented by AROME-MetCoOp. However, the peak values were underestimated in some areas. This indicates that the models underestimate the highest precipitation values; therefore, the results of Paper III (and IV) are more likely to be conservative estimates of possible impact. This should be considered when presenting or applying the findings.

The last part of the modeling chain is the hydrological model, which is the operational flood-forecasting model calibrated and validated using the seNorge dataset. The hydrological model is calibrated on historical observations and may, therefore, not be suitable for calculations of the water balance in a changing climate, when it might be applied to precipitation and temperature values outside the calibration range. In addition, possible changes in land use and vegetation are ignored. One approach to addressing the robustness of a model to changing climate is to calibrate the model with, for example, a cold and dry period and validate it with a warm and wet period. Another approach is to select several parameter sets, and thereby include the uncertainty in parameter estimates. This approach was used in the calibration of the distributed HBV model in Paper IV. For the hydrological model in Paper III and the HBV-lumped model in Paper IV, the intention was to keep the operational model parameters; therefore, it was possible to relate the hydrological output to the flood exceedance levels used operationally. Since the results can be compared to known values and recent events, it can facilitate how the physically

plausible future flood impact is communicated to end users and decision-makers and future challenges disclosed. The atmospheric river events presented in Papers III and IV showed how the future events affected more catchments with higher floods. The findings are supported by climate projection studies that indicate an increase in precipitation and rain-induced floods in the region (e.g., Lawrence and Hisdal 2011; Vormoor et al. 2015). Furthermore, Paper III showed that for selected extreme events, the initial hydrological conditions could alter the ultimate flood level and should therefore be accounted for.

In Paper IV, the added value of using modeling at higher resolution was presented. The peak of the event and, therefore, the value that indicates the highest damage potential were better represented by the hourly distributed HBV model compared to the daily lumped model. Consequently, a high temporal resolution provides even higher specificity, which can be important to how the impact of selected events can be communicated and perceived by, for example, the public, stakeholders, or decision makers. However, results should be communicated carefully as an increased level of detail can give the impression of certainty due to the specificity of the presented event (Sillman et al. 2020). In Paper IV, two catchments were selected to illustrate the effect of changing spatial and temporal resolution. It showed that the future event was smaller than the present for one catchment. This result underlines that events from free-running models will not necessarily affect all catchments within a region. However, in Paper III, the selected extreme precipitation events were well suited to showing the overall impact of atmospheric rivers for a larger region and visualizing the spatial hazard related to these events. The selection criteria for the events are essential and should be targeted to ensure that the data provided are suitable to answer the research question.

### 6.2.1 Outlook

In this setup, the events were chosen based on the 24-h precipitation for a defined selection area. Since precipitation is the most important flood generating variable, this is a natural selection criterion. The high-resolution EC-Earth model provides precipitation data usable as direct input to the hydrological model. An evaluation of a different approach to define the most extreme events could be done by running the hydrological model with EC-Earth

data and select the most extreme events from streamflow data. Backtracking the events will enable an evaluation to whether the events were the same as chosen from the precipitation data, or from similar atmospheric river events, and would provide valuable insight on assessment on local flood impact based on events selected from precipitation data alone.

# Chapter 7

# Conclusions

In the first part of this thesis, the performance of streamflow forecasts applied as combinations of raw, pre-, and postprocessed variables was assessed to answer the following research questions:

*Q1: Will pre- and/or postprocessing schemes improve daily streamflow forecasts, and to which variables should a processing approach be applied?*

The results of Paper I showed that temperature calibration affected the streamflow forecast skill and that preprocessing temperature forecasts could not be neglected in climates with seasonal snow cover. In Paper II, different processing schemes were established by combining the processing of precipitation, temperature, and streamflow. Evaluation of the full dataset showed that all processing schemes performed better than the raw ensemble forecasts. The grid calibration, $P_{cal}$, was the least good preprocessing alternative for the streamflow performance skill. Postprocessing of the raw ensemble forecasts did not perform as well as any combination of pre- and postprocessed ensembles. Preprocessing temperature in combination with postprocessing provided the marginally best performance skill.

*Q2: How do the processing schemes perform under flood conditions, and are there seasonal or regional differences?*

By focusing on the flood dataset, Paper II showed that the processing schemes improved forecast performance also for flood conditions. However, the improvements were not as clear as for the full dataset, and there were larger differences between catchments in skill improvements of the applied processing scheme. Another difference was that the skill of

the forecasts in the flood dataset, independent of the processing scheme, was more sensitive to lead time than that of the forecasts in the full dataset.

Moreover, the results showed that the predictability of floods depends on season. Fewer forecasting schemes were able to predict autumn floods compared to spring floods, especially for long lead times. Some regional differences were clear in the skill of different processing schemes. Coastal areas were better forecasted when preprocessing was included, whereas higher elevations and inland areas showed the highest skill when postprocessing was applied in addition to preprocessing. However, the effect of postprocessing was reduced by lead time. For coastal areas, $P_{bma}$ was of the highest importance for floods and especially for short lead times.

The evaluation of both the full dataset and the flood dataset confirms the hypothesis of this study that processing improves ensemble forecasts. Papers I and II show that all processing schemes improve the forecasts for the diverse Norwegian catchments. However, catchments differ in the optimal processing scheme. For a single catchment, the best processing scheme depends on the dataset used for training the statistical model, season, and lead time. Therefore, if implementing pre- and postprocessing schemes in the operational flood forecasting system in Norway, catchment-specific and adaptable methods should be considered to enable optimized forecasts.

Even though the results clearly show the improvements of applying pre- and postprocessing schemes compared to using the raw ensemble forecasts, there is still work to be done. The catchment specific BMA is a flexible method and showed good results for all variables. However, the results revealed the need for further attention on preprocessing precipitation, especially for high precipitation events. For some events, the precipitation amounts for some individual members were unrealistically high. This might be solved by ensuring that the BMA model is trained on a representative dataset that includes sufficient high precipitation events. Using ECMWF reforecasts might solve this and will be included in future studies.

In the second part of the thesis, a storyline approach was applied to explore the impact of AR-induced floods on western Norway in a future climate. The studies answered the following questions.

*Q3: What is the added value of an event-based storyline approach in climate impact studies?*

The event-based approach provided plausible scenarios in which both the exceedance and magnitude of floods in individual catchments and the total impact over a larger area could be assessed. The findings of this storyline approach support those of previous climate projection studies of Norway indicating more floods at a higher magnitude. Moreover, this thesis was able to evaluate the sensitivity of flood magnitude to initial hydrological conditions by providing initial states with different combinations of soil moisture and snow.

The event-based storyline approach enabled a comparison of the future flood impact to the present flood impact within an operational forecasting chain. The impact assessment could be targeted on specific weather events for a specific location, in this case, the atmospheric rivers affecting western Norway. Furthermore, providing results within a known visual frame is a unique way of presenting plausible future impact in a manner that is easily accessible for end users and decision-makers.

*Q4: What is the added value of higher spatial resolution in climate impact studies?*

The high spatial resolution of the EC-Earth model is well adapted to represent atmospheric rivers and resolve the processes responsible for orographic precipitation on the Norwegian west coast, which is the focus of this thesis. The AROME-MetCoOp model, which is a non-hydrostatic weather forecasting model, gives an even better distribution of precipitation intensities for the complex topography of the Norwegian coast, as well as better spatial representation of temperature. The streamflow magnitudes for the two catchments evaluated show that the better spatial representation of precipitation and temperature by AROME-MetCoOp indicates higher floods compared to the EC-Earth model. Furthermore, implementation of a higher temporal resolution, e.g., by using hourly

data as forcing for the hydrological model, provides better representation of the peak flood, and, consequently, easier assessment of the damage potential of floods.

The storyline approach shows some clear advantages in that the strategically chosen modelling chain makes the results accessible for both scientist and management audience. Models that are validated and well adopted to the purpose provides plausible extreme events. Storylines provide a method to present the impact of individual future flood events for a larger area and is thereby complementary to other climate change studies that for example provide climate factors dimensioning of flood magnitude. Storylines could moreover be used as 'stress test' for how well the society is prepared for future floods. Further extension to the storyline could for example be to include a hydraulic model to evaluate the flood inundated areas, and thereby reveal exposed areas.

# Chapter 8

# References

Azad, R and A. Sorteberg, 2017: Extreme daily precipitation in coastal western Norway and the link to atmospheric rivers. *Journal of Geophysical Research: Atmospheres*, 122, 2080–2095. https://doi.org/10.1002/2016JD025615.

Baatsen, M., Haarsma, R.J., Van Delden, A.J, and de Vries, H., 2015: Severe autumn storms in future Western Europe with a warmer Atlantic Ocean, Clim Dyn. 45(3–4), 949–964. https://doi.org/10.1007/s00382-014-2329-8

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

Beldring, S., Roald, L.A., Engen-Skaugen, T., Førland, E.J., 2006: Climate change impacts on hydrological processes in Norway 2071–2100. *NVE Report No. 5*-2006.

Benedict, I., K. Ødemark, T. Nipen, and R. Moore, 2019: Large-scale flow patterns associated with extreme precipitation and atmospheric rivers over Norway, *Monthly Weather Review*, 147, 1415–1428. https://doi.org/10.1175/MWR-D-18-0362.1

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, 1919–1935. https://doi.org/10.1175/mwr-d-16-0417.1, 2017.

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, SMHI RHO, 7, Norrköping.

Berz, G., Flood disasters: Lessons from the past - Worries for the future, Kassel Reports of Hydraulic Engineering, 9, F1–F9, Herkules, Kassel, Ger., 2000.

Beven, K.., and Binley, A., 1992: The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*. 6 (3): 279–298. doi:10.1002/hyp.3360060305. ISSN 0885-6087.

Bintanja, R., Severijns, C., Haarsma, R., & Hazeleger, W., 2014: The future of Antarctica's surface winds simulated by a high-resolution global climate model: 1. Model description and validation. *Journal of Geophysical Research: Atmospheres*, 119(12), 7136–7159. https://doi.org/10.1002/2013JD020847

Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B., … and Živković, N.,
2017: Changing climate shifts timing of European floods, *Science*, 357(6351), 588–590. https://doi.org/10.1126/science.aan2506.

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., … and Živković, N., 2019: Changing climate both increases and decreases European river floods, *Nature*, 573(7772), 108–111, https://doi.org/10.1038/s41586-019-1495-6.

Bogner, K., Liechti, K., Zappa, M., 2016: Post-processing of stream flows in Switzerland with an emphasis on low flows and floods. *Water*, 8, 115. https://doi.org/10.3390/w8040115

Bogner, K. and Pappenberger, F., 2011: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research*, 47, W07524. https://doi.org/10.1029/2010WR009137

Box, G. E. P. and Cox, D. R., 1964: An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, 26, 211-252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Bremnes, J. B., 2007: Improved calibration of precipitation forecasts using ensemble techniques. Part 2: Statistical calibration methods, *met.no, Report no*. 4, 34 pp., Oslo, Norway, available at: http://met-xpprod.customer.enonic.io/publikasjoner/met-report/met-report-2007 (last accessed: February 1, 2019).

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system, *Monthly Weather Review*., 125, 99–119.       https://doi.org/10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2,

Buizza, R., 2005: EPS skill improvements between 1994 and 2005. *ECMWF Newsletter* No. 104, ECMWF, Shinfield Park, Reading RG2 9AX, UK, 10–14.

Buizza, R., 2010: Horizontal resolution impact on short- and long-range forecast error. *Quarterly Journal of the Royal Meteorological Society*, 136, 1020–1035. https://doi.org/10.1002/qj.613

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M., 2005: Comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems, *Monthly Weather Review*, 133, 1076–1097.   https://doi.org/10.1175/MWR2905.1

Buizza, R., Miller, M., & Palmer, T. N., 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908. https://doi.org/10.1002/qj.49712556006

Clark, M. P., S. Gangopadhyay, L. E. Hay, B. Rajagopalan, and R. L. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*., 5, 243–262. https://doi.org/10.1175/1525-7541(2004)005<0243

Cloke, H. L. and Pappenberger, F. 2009: Ensemble Forecasting: A review. *Journal of Hydrology*, 375(3), 613-626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., ... and Bechtold, P., 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), 553–597.

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z. and Zhu, Y., 2010: Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmosph. Sci. Lett.,* 11: 114-122. https://doi.org/10.1002/asl.261

Dettinger, M., 2011: Climate change, atmospheric rivers, and floods in California – A multimodel analysis of storm frequency and magnitude changes. *Journal of the American Water Resources Association*, 47: 514–523.  https://doi.org/10.1111/j.1752-1688.2011.00546.x

Dixon, K.W., Lanzante, J.R., Nath, M.J., … and Gaitàn, C.F., 2016: Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results? *Climatic Change* 135, 395–408. https://doi.org/10.1007/s10584-016-1598-0

Dominguez, F., Dall'erba, S., Huang, S., Avelino, A., Mehran, A., Hu, H., Schmidt, A., Schick, L., and Lettenmaier, D., 2018: Tracking an atmospheric river in a warmer climate: from water vapor to economic impacts, *Earth System Dynamics*, 9, 249–266, https://doi.org/10.5194/esd-9-249-2018.

Duan, Q., Ajami, Q. H., Gao, X. and 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, 2017: How to pin down uncertainty weather forecasting, https://www.ecmwf.int/en/about/media-centre/news/2017/how-pin-down-uncertainty-weather-forecasting, Figures can be used under the Creative Commons Attribution-Non-Commercial-No-Derivatives-4.0-Unported Licence: https://www.ecmwf.int/en/terms-use. Retrieved on September 26, 2020.

ECMWF, 2020a: Forecaster User Guide: Model Data Assimilation. https://confluence.ecmwf.int/display/FUG/2.5+Model+Data+Assimilation%2C+4D-Var. Retrieved on October 10, 2020

ECMWF, 2020b: Forecasts: Documentation and Support https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model, Retrieved on May 31, 2020

ECMWF, 2020c: Change in Model versions: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model, Retrived on October 10, 2020

EFAS, 2014: European Flood Awareness Senter, *EFAS Bulletin* 2014[6], ECMWF, Reading, UK https://www.efas.eu/sites/default/files/efasBulletins/2014/bulletin_oct-nov_14.pdf, retrieved on July 21, 2020

Engeland, K., Steinsland, I., Johansen, S.S., and Petersen-Øverleir, A., 2016: Effects of uncertainties in hydrological modelling. A case study of a mountainous catchment in Southern Norway, *Journal of Hydrology*, 536, 147-160. Doi: 10.1016/j.jhydrol.2016.02.036

Espinoza, V., Waliser, D. E., Guan, B., Lavers, D. A., & Ralph, F. M., 2018: Global analysis of climate change projection effects on atmospheric rivers. *Geophysical Research Letters*, 45, 4299–4308. https://doi.org/10.1029/2017GL076968

Eum, H.-I. and Cannon, A.J., 2017: Intercomparison of projected changes in climate extremes for South Korea: application of trend preserving statistical downscaling methods to the CMIP5 ensemble. *International Journal of Climatology*., 37: 3381–3397. https://doi.org/10.1002/joc.4924

Felder, G., A. Zischg and R. Weingartner, 2017: The effect of coupling hydrologic and hydrodynamic models on probable maximum flood estimation, *Journal of Hydrology*, 550, 157–165. https://doi.org/10.1016/j.jhydrol.2017.04.052

Follestad, T., and G. Høst, 1998: A Statistical model for the uncertainty in meteorological forecasts, with applications to the Knappom and Røykenes catchments, *NVE HYDRA notat* 12/98, pp. 15, Oslo, Norway. http://publikasjoner.nve.no/hydra/notat/1998/hydranotat1998_12.pdf

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. https://doi.org/10.1175/2009MWR3046.1

Friederichs, P., & Thorarinsdottir, T. L., 2012: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7), 579–594. https://doi.org/10.1002/env.2176

Frogner, I-L., Singleton, AT, Køltzow, MØ, Andrae, U. Convection-permitting ensembles, 2019a: 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.,…, and 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.

Fundel, F. and Zappa, M., 2011: Hydrologic ensemble forecasting in mesoscale catchments: Sensitivity to initial conditions and value of reforecasts. *Water Resources Research*, 47, W09520. https://doi.org/10.1029/2010WR009996

Gneiting, T., Raftery, A. E., Westveld III, A.H., and Goldman, T., 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5), 1098–1118. https://doi.org/10.1175/MWR2904.1

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

Gottschalk, L., Jensen, J. L., Lundquist, D., Solantie, R., and Tollan, A., 1979: Hydrologic regions in the Nordic countries, *Hydrology Research*., 10, 273–286. https://doi.org/10.2166/nh.1979.0010

Haarsma, R. J., Hazeleger, W., Severijns, C., Vries, H., Sterl, A., Bintanja, R., Oldenborgh, G. J., Brink, H.W., 2013: More hurricanes to hit Western Europe due to global warming. *Geophysical Research Letters*. 40(9) 1783–1788. https://doi.org/10.1002/grl.50360

Haiden, T., Janousek, M., Bidlot, J., Buizza, R., Ferranti, L., Prates, F., and Vitart, F., 2018: Evaluation of ECMWF forecasts, including the 2018 upgrade. European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, England.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Hamill, T. M., and Colucci, S. J., 1997: Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, *125*(6), 1312–1327. https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2

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

Hanssen-Bauer, I & EJ Førland (1998): Annual and seasonal precipitation variations in Norway 1896–1997. *Klima-Report* 27/98, Norwegian Meteorological Institute.

Hanssen-Bauer, I and Ø Nordli, 1998: Annual and seasonal temperature variations in Norway 1876– 1997. *Klima-Report* 25/98, Norwegian Meteorological Institute.

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. 2017: Climate in Norway 2100 - a knowledge base for climate adaption. *Technical Report* 1, Norwegian Climate Service Centre.

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., … and Willèn U., 2010: EC-Earth: a seamless earth-system prediction approach in action. *Bulletin of the American Meteorological Society*, *91*(10), 1357-1364. https://doi.org/10.1175/2010BAMS2877.1

Hazeleger, W., van den Hurk, B.J.J.M., Min, E., van Oldenborgh, G.J., Petersen, A.C., Stainforth, D.A., Vasileiadou, E., and Smith, L. A, 2015: Tales of future weather, *Nature Climate Change*, 5, 107–113. https://doi.org/10.1038/NCLIMATE2450

Hawkins, E., & Sutton, R., 2009: The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Hegge, K., 1968: Flomvarsling i Glomma, *Teknisk Ukeblad, 9* (in Norwegian), Oslo Norway.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2, 2000.

IPCC, 2010: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)]. *IPCC Working Group I Technical Support Unit*, University of Bern, Bern, Switzerland, pp. 117.

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., ..., and Georgopoulou, E., 2014: EURO-CORDEX: new high-resolution climate change projections for European impact research. *Regional environmental change*, 14(2), 563-578. https://doi.org/10.1007/s10113-013-0499-2

Jacobeit J, Glaser R, Luterbacher J, Nonnenmacher M, Wanner H, 2003: Links between flood events in Central Europe since AD 1500 and the large-scale atmospheric circulation. *Geophysical Research Letter* 30:1172. https://doi.org/10.1029/2002GL016433

Ji, L., X. Zhi, S. Zhu, and K. Fraedrich, 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., and Primo, C., 2008: Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136(6), 2133–2139. https://doi.org/10.1175/2007MWR2219.1

Jolliffe, I. T., and Stephenson, D. B. (Eds.)., 2012: Forecast verification: a practitioner's guide in atmospheric science, Oxford, UK, John Wiley & Sons.

Kalnay, E., 2003: Atmospheric modeling, data assimilation and predictability, New York, USA, Cambridge University Press. 7[th]ed 2012.

Kavetski, D, G. Kuczera, S.W. Franks, 2006: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory *Water Resources Research*, 42, pp. 1–9. https://doi.org/10.1029/2005WR004368

Kay, J. K, H. M. Kim, Y.-Y. Park, and J.Son, 2013: Effect of doubling the ensemble size on the performance of ensemble prediction in the warm season using MOGREPS implemented at the KMA, *Advances in Atmospheric Sciences*. 30, 1287–1302. https://doi.org/10.1007/s00376-012-2083-y

Klein Tank, A. M. G., and G. P. Können, 2003: Trends in indices of daily temperature and precipitation extremes in Europe, 1946–99. *Journal of Climate*, 16, 3665–3680. https://doi.org/10.1175/1520-0442(2003)016<3665:TIIODT>2.0.CO;2.

Krzysztofowicz, R., 2001: The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249, 2–9. https://doi.org/10.1016/S0022-1694(01)00420-6

Langsrud, Ø., Frigessi, A., and Høst, G., 1998: Pure model error for the HBV model, NVE, *HYDRA notat 4/98*, 28 pp., Oslo, Norway.
http://publikasjoner.nve.no/hydra/notat/1998/hydranotat1998_04.pdf

Langsrud, Ø., Høst, G., Follestad, T., Frigessi, A., and Hirst, D., 1999: Quantifying uncertainty in HBV runoff forecasts by stochastic simulations, NVE, *HYDRA notat 2/99*, 38 pp., Oslo, Norway.
http://publikasjoner.nve.no/hydra/notat/1999/hydranotat1999_02.pdf

Lavers, D.A., and G. Villarini, 2013: The nexus between atmospheric rivers and extreme precipitation across Europe, *Geophysical Research Letters*, 40(12), 3259–3264.
https://doi.org/10.1002/grl.50636

Lavers, D.A., and G. Villarini, 2015: The contribution of atmospheric rivers to precipitation in Europe and the United States, *Journal of Hydrology*, **522**, 382–390.
https://doi.org/10.1016/j.jhydrol.2014.12.010

Lawrence, D., and Hisdal, H., 2011: Hydrological projections for floods in Norway under a future climate, Norwegian Water Resources and Energy Directorate, *NVE Report 2011*:5, ISBN: 978-82-410-0753-8

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, *102*(6), 409-418. ttps://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2

Leutbecher, M., & Palmer, T. N., 2008: Ensemble forecasting. *Journal of computational physics*, *227*(7), 3515-3539. https://doi.org/10.1016/j.jcp.2007.02.014

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., et al., 2017: Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Quarterly Journal of the Royal Meteorological Society,* 143: 2315–2339.  https://doi.org/10.1002/qj.3094

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z., 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting, *WIRES Water*, 4, e1246. https://doi.org/10.1002/wat2.1246.

Liu, J. G., and Z. H. Xie, 2014: BMA probabilistic quantitative precipitation forecasting over the Huaihe basin using TIGGE multimodel ensemble forecasts. *Monthly Weather Review*, 142, 1542–1555. https://doi.org/10.1175/MWR-D-13-00031.1

Lorenz, E. N., 1963: Section of planetary sciences: The predictability of hydrodynamic low. *Transactions of the New York Academy of Sciences*, *25*(4 Series II), 409–432.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3), 289–307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., ... & Raper, S. C. B., 2007: Global climate projections, in Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the *Intergovernmental Panel on Climate Change*. [edited by S. Solomon et al.], pp. 747–846, Cambridge University Press, Cambridge, UK. https://www.ipcc.ch/report/ar4/wg1/

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. D. J., Boer, G., Danabasoglu, G. N., … and Stockdale, T., 2009: Decadal Prediction: Can It Be Skillful?. *Bulletin of American Meteorological Society*, 90, 1467–1486, https://doi.org/10.1175/2009BAMS2778.1.

Miller M., R. Buizza, J. Haseler, M. Hortal, P. Janessen and A. Untch, 2010: Increased resolution in the ECMWF deterministic and ensemble prediction systems, *ECMWF Newsletter* No. 124 pp.10–16.

Mohr, M., 2008: New routines for gridding of temperature and precipitation observations for "seNorge.no", met.no, *Note n*o. 8, 40 pp., Oslo, Norway, 2008.

Monhart, S., Zappa, M., Spirig, C., Schär, C., and Bogner, K., 2019: Subseasonal hydrometeorological ensemble predictions in small- and medium-sized mountainous catchments: benefits of the NWP approach, *Hydrol. Earth Syst. Sci.*, 23, 493–513, https://doi.org/10.5194/hess-23-493-2019.

Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T., 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, *122*(529), 73-119. https://doi.org/10.1002/qj.49712252905

Müller, M., Homleid, M., Ivarsson, K.-I., Køltzow, M. A. Ø., Lindskog, M., Midtbø, K. H., . . . Vignes, O., 2017: AROME-MetCoOp: A Nordic Convective-Scale Operational Weather Prediction Model. *Weather and Forecasting*, 32(2), 609–627. https://doi.org/10.1175/waf-d-16-0099.1

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

NCCS, 2020: Norwegian Centre for Climate Services (NCCS), data from NCCS and figures kindly produced by dr. W.K. Wong.

https://klimaservicesenter.no/faces/desktop/article.xhtml?uri=klimaservicesenteret/om-norsk-klimaservicesenter.

Neiman, P.J., A.B. White, F.M. Ralph, D.J. Gottas, and S.I. Gutman, 2009: A Water Vapor Flux Tool for Precipitation Forecasting. U.K. *Water Management Journal* (Special Issue on Weather Radar for Water Management) 162:83-94. https://doi.org/10.1680/wama.2009.162.2.83

Nieto, R., Durán-Quesada, A. M., Drumond, A., Yu, L., Yoshimura, K., Dominguez, F., Trigo, R. M., Stohl, A., and Gimeno, L., 2012: Oceanic and terrestrial sources of continental precipitation, *Review of Geophysics*, 50, RG4003.  https://doi.org/10.1029/2012RG000389.

NOU-16, 1996: Norges offentlige utredninger (NOU)-16: Tiltak mot flom ( Norwegian public reports: Measures against floods) https://www.regjeringen.no/no/dokumenter/NOU-1996-16/id140631/?ch=4#kap7, Only available in Norwegian


Owens, R. G. and Hewson, T. D., 2018: ECMWF Forecast User Guide, ECMWF, Reading, UK. https://doi.org/10.21957/m1cs7h.

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G. J., Steinheimer M, & Weisheimer, A., 2009: Stochastic parametrization and model uncertainty. ECMWF Research Department *Technical Memorandum. 598, pp. 42.*

Parrish, M. A., Moradkhani, H., & DeChant, C. M., 2012: Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. *Water Resources Research*, 48(3), W03519. https://doi.org/10.1029/2011WR011116.

Persson, A., 2015: User guide to ECMWF forecast products, edited by: Andersson, E. and Tsonevsky, I., ECMWF, Reading, UK, pp 129.

Ralph, F. M., and Dettinger, M. D., 2011: Storms, floods, and the science of atmospheric rivers. Eos, Trans., *American Geophysical Union*, 92(32), 265–266. *https://doi.org/*10.1029/2011EO320001

Ralph, F. M., Dettinger, M., Lavers, D., Gorodetskaya, I. V., Martin, A., Viale, M., . . . Spackman, J. R., 2017: Atmospheric rivers emerge as a global science and applications focus. *Bulletin of the American Meteorological Society*, 98(9), 1969–1973. https://doi.org/10.1175/BAMS-D-16-0262.1

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M., 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Refsgaard, J.C., and B. Storm, 1996: Construction, calibration and validation of hydrological models M.B. Abbott, J.C. Refsgaard (Eds.), *Distributed Hydrological Modelling*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 41–54. https://doi.org/10.1007/978-94-009-0257-2_3

Reggiani, P., Renner, M., Weerts, A. H. and van Gelder, P. A. H. J. M., 2009: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, 45, W02428. https://doi.org/10.1029/2007WR006758

Reitan, T., Petersen-Øverleir, 2009: A. Bayesian methods for estimating multi-segment discharge rating curves. *Stoch Environ Res Risk Assess* **23,** 627–642. https://doi.org/10.1007/s00477-008-0248-0

Roald L. A, 2013 [Book], Flom i Norge, Vestfossen, Norway, Tom & Tom, ISBN 8292916156, 9788292916155

Roald, L., A., 2008: Rainfall Floods and Weather Patterns, NVE Consultancy Report, 14-2018, pp. 44, http://publikasjoner.nve.no/oppdragsrapportA/2008/oppdragsrapportA2008_14.pdf

Roald, L.A., Beldring, S., Engen Skaugen T., Førland, E., Benestad, R., 2006: Climate change impacts on streamflow in Norway. *NVE Oppdragsrapport* A nr 1-2006, 73 pp.

Rodwell, M. J., 2005/2006: Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better. *ECMWF Newsletter* No. 106, 17–22, ECMWF, Shinfield Park, Reading RG2 9AX, UK, 10–14. https://www.ecmwf.int/sites/default/files/elibrary/2005/14619-newsletter-no106-winter-200506.pdf (last accessed October 15, 2020).

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 *NVE-Publication*, 7, 1–26.

Sælthun, N.R., Aittoniemi, P., Bergström, S., Einarsson, K., Jóhannesson, T., Lindström, G., Ohlsson, P.-E., Thomsen, T., Vehviläinen, B. and Aamodt, K.O., 1998: Climate Change Impacts on Runoff and Hydropower in the Nordic Countries. Final report from the project "Climate Change and Energy Production". *TemaNord 1998*, 552, Nordisk ministerråd, København, 170 s.

Sælthun, N.R., Bogen, J., Flood, M.H., Laumann, T., Roald, L.A., Tvede, A.M., and Wold. B., 1990: Klimaendringer og vannressurser. Bidrag til Den interdepartementale klimautredningen. *NVE Publikasjon V* 30, 104 s.

Schaller, N., A. L. Kay, R. Lamb, N. R. Massey, G. J. van Oldenborgh, F. E. L. Otto, S. N., Sparrow, R. Vautard, P. Yiou, I. Ashpole, A. Bowery, S. M. Crooks, K. Haustein, C. Huntingford, W. J. Ingram, R. G. Jones, T. Legg, J. Miller, J. Skeggs, D. Wallom, A. Weisheimer, S. Wilson, P. A. Stott and M. R. Allen, 2016. Human influence on climate in the 2014 Southern England winter floods and their impacts. *Nature Climate Change*, 6(6), 627–634. https://doi.org/10.1038/nclimate2927

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

Seierstad, I., Kristiansen, J., and Nipen, T., 2016: Better temperature forecasts along the Norwegain coast, *ECMWF newsletter*, 148, available at: https://www.ecmwf.int/en/newsletter/148/news/better-temperature-forecasts-along-norwegian-coast (last accessed July 21, 2020).

Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mendoza, P., and 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, 2018.

Shepherd, T.G., Boyd, E., Calel, R.A. et al., 2018: Storylines: an alternative approach to representing uncertainty in physical aspects of climate change, *Climate Change*, 151 (3-4), p 555-57, https://doi.org/10.1007/s10584-018-2317-9

Shepherd, T. G., 2019: Storyline approach to the construction of regional climate change information, *Proceedings of the Royal Society A*, 475:20190013, https://doi.org/10.1098/rspa.2019.0013

Sheridan, P., Smith, S., Brown, A. and Vosper, S., 2010: A simple height-based correction for temperature downscaling in complex terrain. *Meteorological Applications*, 17: 329–339.  https://doi.org/10.1002/met.177

Sillmann, J, Kharin, Zwiers, F. W., Zhang, X., Bronaugh, D., 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *Journal of Geophysical Research: Atmospheres,* 118 (6), 2473–2493. https://doi.org/10.1002/jgrd.50188.

Sillmann, J., Shepherd, T., van den Hurk, B., Hazeleger, W., Martius-Rompainen, O., Zscheischler, J., 2019: Physical modeling supporting a storyline approach. *CICERO*

*Policy Note 2019*:01, available at: https://www.cicero.oslo.no/en/publications-and-events_twex/workshop-on-physical-modeling-supporting-a-storyline-approach.

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. https://doi.org/10.1175/MWR3441.1

Sloughter, J. M., Gneiting, T., & Raftery, A. E., 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, *105*(489), 25–35. https://doi.org/10.1198/jasa.2009.ap08615

Sorteberg, A., Lawrence, D., Dyrrdal, A.V., Mayer, S., Engeland, K. (editors), 2018: Climate changes in short duration extreme precipitation and rapid onset flooding – implications for design values, Norwegian Center for Climate Services, *NCCS Report 2018*:1, https://cms.met.no/site/2/klimaservicesenteret/rapporter-og-publikasjoner/_attachment/13537?_ts=163df95ff7b

Stohl, A., Forster, C., and 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

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bullentine of the American Meteorological Society*, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1

Thiboult, A., Anctil, F., and Boucher, M.-A., 2016: Accounting for three sources of uncertainty in ensemble hydrological forecasting, *Hydrology and Earth System Sciences*, 20, 1809–1825. https://doi.org/10.5194/hess-20-1809-2016.

Toth, Z., & Kalnay, E., 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, *74*(12), 2317-2330. https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2

Tveito, O. E., Bjørdal, I., Skjelvåg, A. O., and 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.

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

Van der Linden, E. C., Haarsma, R. J., and van der Schrier, G., 2018: Resolution-dependence of future European soil moisture droughts. *Hydrology and Earth System*

*Sciences Discussion*, 2018, 1–31. https://doi.org/10.5194/hess-2018-226

Van Haren, R., Haarsma, R. J., De Vries, H., Van Oldenborgh, G. J., & Hazeleger, W., 2015: Resolution dependence of circulation forced future central European summer drying. *Environmental Research Letters*, 10(5), 055002. https://doi.org/10.1088/1748-9326/10/5/055002

Vannitsem, S., Wilks, D. S., and Messner, J. W. (Eds.), 2018: Statistical Postprocessing of Ensemble Forecasts, Elsevier, ISBN 9780128123720. https://doi.org/10.1016/B978-0-12-812372-0.09988-X.

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A.,Hibbard, K., Hurtt, G. C., Kram,T., Krey, V., Lamarque, J-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., Rose, S.K., 2011: The representative concentration pathways: an overview. Clim. Change 109, 5-31. https://doi.org/10.1007/s10584-011-0148-z

Verkade, J. S., Brown, J. D., Reggiani, P., and 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., Heistermann, M., and Bronstert, A., 2015: Climate change impacts on the seasonality and generation processes of floods – projections and uncertainties for catchments with mixed snowmelt/rainfall regimes, *Hydrology and Earth System Sciences*., 19, 913–931. https://doi.org/10.5194/hess-19-913-2015.

Vormoor, K., Lawrence, D., Schlichting, L., Wilson, D., and 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. https://doi.org/10.1016/j.jhydrol.2016.03.066

Vrugt, J.A., Gupta, H.V., Bouten, W. and Sorooshian, S., 2003: A Shuffled Complex Evolution Metropolis Algorithm for Optimization and Uncertainty Assessment of Hydrological Model Parameters, *Water Resour. Res.*, 39, 1201, doi:10.1029/2002WR001642

Vrugt, J. A., & Robinson, B. A., 2007: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, 43, W01411. https://doi.org/10.1029/2005WR004838

Wetterhall, F., Pappenberger, F., Alfieri, L., Cloke, H.L., Thielen-del Pozo, J., Balabanova, S., Daňhelka, J., Vogelbacher, A., Salamon, P., and 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

Whan, K., Sillmann, J., Schaller, N., and Haarsma, R., 2020: Future changes in atmospheric rivers and extreme precipitation in Norway. *Climate Dynamics*, 54(3), 2071–2084. https://doi.org/10.1007/s00382-019-05099-z

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16(3), 361–368. https://doi.org/10.1002/met.134

Wilks, D. S., and Hamill, T. M., 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, 135(6), 2379–2390. https://doi.org/10.1175/MWR3402.1

Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R., 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging, *Monthly Weather Review*, 135, 1364–1385. https://doi.org/10.1175/MWR3347.1.

Wolff, M. A., Isaksen, K., Petersen-Øverleir, A., Ødemark, K., Reitan, T., and Brækkan, R., 2015: Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: results of a Norwegian field study, *Hydrol. Earth Syst. Sci.,* 19, 951–967, https://doi.org/10.5194/hess-19-951-2015.

Wu, L., Zhang, Y., Adams, T., Lee, H., Liu, Y., and Schaake, J., 2018: Comparative evaluation of three Schaake shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. *Journal of Hydrometeorology*, *19*(3), 575–598. https://doi.org/10.1175/JHM-D-17-0054.1

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. https://doi.org/10.1016/j.jhydrol.2019.124002

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, and Fundel F., 2011: Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research* 100: 246–262.  https://doi.org/10.1016/j.atmosres.2010.12.005

Zappa, G., L. C. Shaffrey, and K. I. Hodges, 2013: The ability of CMIP5 models to simulate North Atlantic extratropical cyclones. *Journal of Climate* 26(15), 5379–5396. https://doi.org/10.1175/JCLI-D-12-00501.1

Zhu, Y., and Newell, R. E., 1998: A proposed algorithm for moisture fluxes from atmospheric rivers. *Monthly Weather Review*, 126(3), 725–735. https://doi.org/10.1175/1520-0493(1998)126<0725:apafmf>2.0.co;2

# Chapter 9

# Papers

# Paper I:

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

**T. J. Hegdahl, K. Engeland, I. Steinsland, and L. M. Tallaksen**

Hydrology and
Earth System
Sciences

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

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

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

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

**Abstract.** In this study, we used meteorological ensemble forecasts as input to hydrological models to quantify the uncertainty in forecasted streamflow, with a particular focus on the effect of temperature forecast calibration on the streamflow ensemble forecast skill. In catchments with seasonal snow cover, snowmelt is an important flood-generating process. Hence, high-quality air temperature data are important to accurately forecast streamflows. The sensitivity of streamflow ensemble forecasts to the calibration of temperature ensemble forecasts was investigated using ensemble forecasts of temperature from the European Centre for Medium-Range Weather Forecasts (ECMWF) covering a period of nearly 3 years, from 1 March 2013 to 31 December 2015. To improve the skill and reduce biases of the temperature ensembles, the Norwegian Meteorological Institute (MET Norway) provided parameters for ensemble calibration, derived using a standard quantile mapping method where HIRLAM, a high-resolution regional weather prediction model, was used as reference. A lumped HBV (Hydrologiska Byråns Vattenbalansavdelning) model, distributed on 10 elevation zones, was used to estimate the streamflow. The results show that temperature ensemble calibration affected 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 affected by subzero temperature improvements, which accounted for the biggest temperature biases and corrections during autumn and winter. The skill differs between regions. In particular, there is a cold bias in the forecasted temperature during autumn and winter along the coast, enabling a large improvement by calibration. The forecast skill was partly related to elevation differences and catchment area. Overall, it is evident that temperature forecasts are important for streamflow forecasts in climates with seasonal snow cover.

## 1 Introduction

Floods can severely damage infrastructure, buildings, and farmland, and can 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 components, all affected 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 Vattenbalansavdelning model (HBV) (Bergström, 1976; Sælthun, 1996; Beldring, 2008) is used as the hydrological forecasting model, which combined with statistical uncertainty models (Langsrud et al., 1998, 1999) provides probabilistic streamflow forecasts. The uncertainty model accounts for the strong autocorrelation in forecast errors and estimates an uncertainty band around the deterministic temperature, precipitation, and streamflow forecasts.

An alternative approach to estimate probabilistic streamflow forecasts is to use meteorological ensemble forecasts from numerical weather prediction models as a means to account for uncertainty in the forcing. The meteorological ensemble forecasts are created by perturbing both the initial states and the physics tendencies 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). HEPSs 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 modeling 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 can be calibrated by applying statistical techniques correcting bias and under- or over-dispersion (Hamill and Colucci, 1997; Persson, 2015). Examples of methods used to calibrate meteorological ensembles include 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 (Gneiting et al., 2005; Wilks and Hamill, 2007), 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 in how the spread and bias are corrected. Preprocessing (from a hydrological perspective) refers to all techniques used to change the output from a meteorological model and includes calibration (described above) and downscaling. Downscaling implies resampling from the original forecast grid size to a grid of higher resolution, and both statistical (e.g., interpolation) and dynamical (e.g., a regional weather forecast model) techniques can be used (Schaake et al., 2010). A recent review of preprocessing methods is given in Li et al. (2017) and the textbook edited by Vannitsem et al. (2018).

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 nonlinear since streamflow depends on temperature thresholds for rain–snow partitioning and for snowmelt and freeze processes. The snowmelt and freeze processes depend on the state of the system; i.e., snow is needed to generate snowmelt. For temperatures well below $0\,^\circ\text{C}$, the streamflow is not sensitive to temperature, whereas for temperatures around $0\,^\circ\text{C}$ relatively small changes in temperature might control if precipitation falls as rain or snow, and consequently whether streamflow is generated or not. Most Norwegian catchments experience a seasonal snow cover, but are otherwise diverse in terms of the length of the snow season and topographic complexity (Rizzi et al., 2017).

Downscaling and interpolating air temperature in complex topography are both 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; Sheridan 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, found 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 isolated effect of the uncertainties in temperature forecasts has not yet been 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 questions:

– 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 temperature and streamflow 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 preprocessing setup from MET Norway to 139 catchments in Norway. Three years of operational ECMWF forecasts from 2013 to 2015 were used to regenerate streamflow forecasts, and the skill of temperature and streamflow forecasts was systematically evaluated for these catchments. To investigate the isolated effect of the temperature ensembles on the streamflow forecasts, the observed seNorge precipitation (Tveito et al., 2005) was used instead of the precipitation ensemble forecasts to run the hydrological model. Finally, a case study 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 spatial variations in climate and topography, and a recent overview of 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 yr$^{-1}$), 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 Köppen–Geiger system; see Peel et al., 2007). The mean annual runoff 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 subzero 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 is 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 15 447 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 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), middle (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 hydroclimatological zones. Rainfall floods dominate in the south, west, and middle, whereas snowmelt floods dominate in the 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 (E), Foennerdalsvatn (W), and Viksvatn (W) were used to visualize the challenges in temperature forecasts, and both uncalibrated and calibrated ensemble values will be presented for these three catchments. Viksvatn (W) and Foennerdalsvatn (W) are located in western Norway and are both catchments with some glaciers (~ 3 % and 47 %, respectively). Gjuvaa (E) is non-glaciered and located inland (Fig. 1, left). The Bulken (W) catchment was chosen to demonstrate the effect of temperature calibration on the streamflow forecast for a snowmelt-driven flood event.

### 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 km × 1 km) product, referred to as seNorge (available at http://www.seNorge.no/, last access: 1 February 2019, Tveito et al., 2005). In this study, we used version 1.1. For this version, gridded temperature is calculated by kriging, 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 km × 1 km grid. A constant gradient of 10 % per 100 m beneath 1000 m above sea level (m a.s.l.) and 5 % per 100 m above 1000 m a.s.l. is applied to account for elevation gradients in precipitation (details can be found in Tveito,
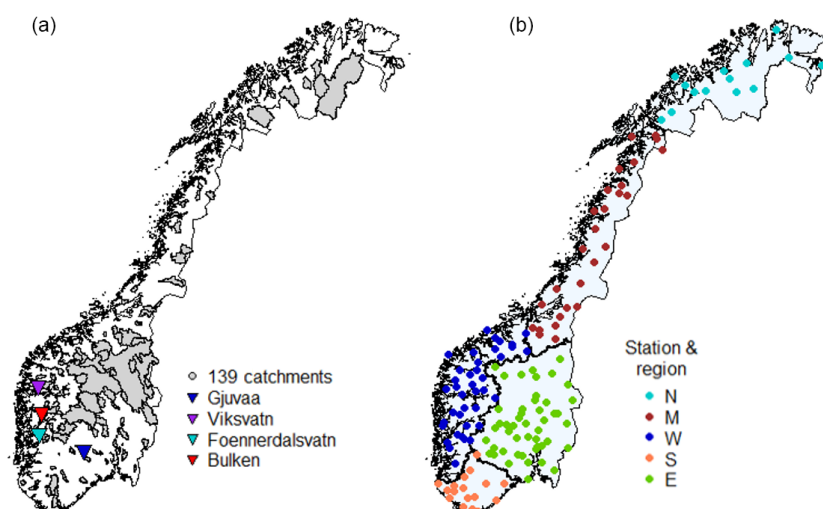
**Figure 1.** The maps for Norway indicate the 139 catchments used in this study. Panel **(a)** shows the catchment boundaries including the location of four selected catchments. Please note that many catchments are relatively small and difficult to detect. The locations of the catchment gauging stations are shown in **(b)**. Norway was grouped into five regions (N: north, M: middle, W: west, S: south, and E: east); all regions are marked with different colors and regional boundaries.

2002; Tveito et al., 2005; and Mohr, 2008). The seNorge data are available from 1 January 1957, and in this study we used data for the period 1 March 2013 to 31 December 2015 in the forecasting mode and 1 January 1958 to 31 December 2012 to calculate the temperature and streamflow climatology (Sect. 3.2). The seNorge precipitation substitutes the precipitation forecasts in the ensemble forecasting chain, and hence the isolated effect of temperature calibration on streamflow forecasts was obtained. We hereby denote seNorge temperature and precipitation, $T_{o[lat,long,t]}$ and $P_{o[lat,long,t]}$, respectively, where $t$ is an index for observation time. Latitude (lat) and longitude (long) represent the grid indexing.

### 2.2.2 Hydrological model – HBV

The HBV model (Bergström, 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. The model uses catchment average temperature and precipitation as input. Each catchment is divided into 10 elevation zones, each covering 10 % of the total catchment area. The catchment average precipitation and temperature are elevation adjusted to each elevation zone using catchment-specific lapse rates. In this study, we used the operational model setup which has been calibrated for each catchment individually. PEST, a software for parameter estimation and uncertainty analysis (Doherty, 2015), was used to optimize the HBV parameters, with the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) and volume bias as calibration metrics. The calibration period,

1996–2012, gives a mean Nash–Sutcliffe efficiency value of 0.77, with zero volume bias for the 139 catchments. The validation period, 1980–1995, shows a mean Nash–Sutcliffe efficiency value of 0.73, with a mean volume bias of 5 % (Ruan Gusong, personal communication, 15 June 2016). We used one optimal parameter set for each catchment and therefore ignored uncertainty arising from parameter estimation and the hydrological model.

### 2.2.3 Reference streamflow

Reference 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, reference streamflow was used as a benchmark 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 reference streamflow.

### 2.2.4 Temperature ensemble forecasts

We used the ECMWF temperature forecast ensemble (ENS) for the period 1 March 2013 to 31 December 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 Sect. 3.1.2) plus spring 2013 (cycle 38r1/2) in order to include one more snowmelt season. In short, 50 ensemble members of ENS
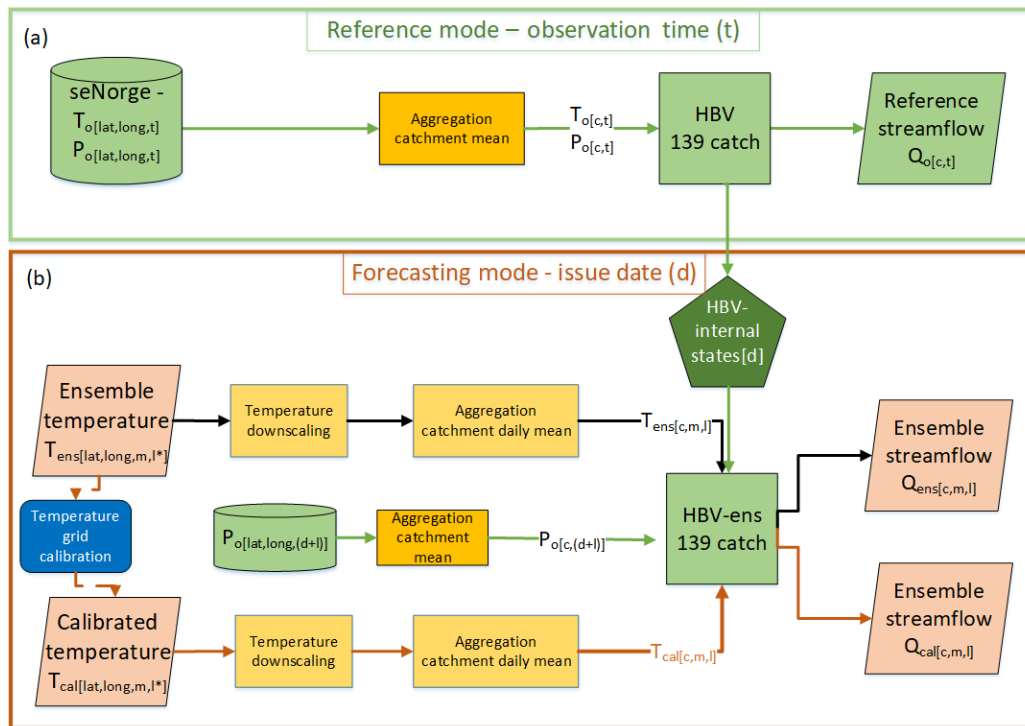
**Figure 2.** Conceptual diagram of the ensemble forecasting chain. Panel **(a)** shows the reference mode that is the calculation of reference streamflow using the HBV model with catchment aggregated daily mean values of seNorge temperature ($T_o$) and precipitation ($P_o$). In the forecasting mode **(b)**, ECMWF temperature ensembles are downscaled to $1\,\text{km} \times 1\,\text{km}$ prior to catchment aggregation. Calibrated temperature ($T_{cal}$) is estimated from $T_{ens}$, applying a grid calibration at $5\,\text{km} \times 5\,\text{km}$ resolution. Daily average forecast values ($T_{ens}$ or $T_{cal}$) and observed precipitation ($P_o$) are used to force the hydrological model at the forecasting issue date ($d$), with internal states from the reference mode.

are generated by adding small perturbations to the forecast initial conditions and model physics schemes, subsequently running the model with different perturbed conditions. The ensemble represents the temperature forecast uncertainty. A more detailed description of the ECMWF ENS system is provided in, for example, Buizza et al. (1999, 2005) and Persson (2015). For each issue date $d$, 51 ensemble members $T_{\text{ens}[\text{lat},\text{long},m,l^*]}$ are provided for a lead time of up to 246 h, where $m$ is the ensemble member and $l^*$ the lead time in 6 h intervals. In this study, we used the forecasts issued at 00:00 UTC and aggregated daily values for the meteorological 24 h period defined as 06:00–06:00 to provide forecasts for lead times of up to 9 days. The observational time $t$ for a forecast is $d + l^*$.

## 3 Methods

### 3.1 Ensemble forecasting chain

Figure 2 shows the forecasting modeling chain designed for this study. The green frame presents the observational reference mode that determines the internal states for the forecasting issue date, $d$, in the red frame. This reference

mode was also used to estimate reference streamflow $Q_{\text{o}[c,t]}$ (see Sect. 2.2.3). The seNorge temperature and precipitation ($T_{\text{o}[c,t]}$ and $P_{\text{o}[c,t]}$), aggregated to each catchment $c$, were used to force the hydrological model in the observational reference mode. The red frame illustrates the forecasting mode, including the post-processing of temperature forecasts. The hydrological ensemble forecasts were estimated using downscaled raw temperature ensemble forecasts ($T_{\text{ens}[c,m,l]}$; see Sect. 3.1.1) or downscaled and calibrated temperature ensemble forecasts ($T_{\text{cal}[c,m,l]}$; see Sect. 3.1.2) and observed precipitation ($P_{\text{o}[c,d+1]}$) as forcing, where $m$ is the ensemble member and $l$ is lead time in days. All temperature forecasts were aggregated to daily time steps since the operational HBV model runs on a daily time step and the seNorge data used as a reference provide only daily values. In the forecasting mode, each temperature ensemble member was used as input and run as a separate deterministic forecast. All hydrological forecasts were estimated for all nine lead times. Note that for each issue date $d$, the same internal states of the HBV model were used for all ensemble member runs. Thus two sets of streamflow ensemble forecasts ($Q_{\text{ens}[c,m,l]}$ and $Q_{\text{cal}[c,m,l]}$) 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

In this paper the term downscaling refers to 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, which represent the average temperature for the grid cell, was interpolated from a horizontal resolution of $0.25°$ ($\sim 30$ km) to the $1$ km $\times 1$ km seNorge grid, using the nearest-neighbor method and aggregated to daily values to match the spatial and temporal resolution of the seNorge data. Due to elevation differences between the ECMWF and seNorge grid elevations, we corrected the ensemble temperature at the $1$ km $\times 1$ km scale by applying a standard atmospheric lapse rate of $-0.65$ °C $100$ m$^{-1}$. 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 et al., 2016; Bremnes, 2007) 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. This grid calibration was used in the operational post-processing chain for meteorological forecasts including the forecasts published on http://yr.no (last access: 1 February 2019). MET Norway uses HIRLAM (High Resolution Limited Area Model, Bengtsson et al., 2017) temperature forecasts (on a $4$ km $\times 4$ km grid) to provide a reference for parameter estimation (calibration). HIRLAM is suitable as a reference since it provides a continuous field covering all of Norway at a sub-daily time step. In addition, HIRLAM gives higher skill and is less biased than ENS (Engdahl and Homleid, 2015). To establish the calibration parameters, MET Norway used both ENS reforecasts (Owens and Hewson, 2018) and HIRLAM data from July 2006 to December 2011 interpolated to a $5$ km $\times 5$ km grid. The ENS reforecast is a five-member ensemble generated from the same model cycle (40r1 and 41r1) as ENS. For each grid cell, monthly unique quantile transformation coefficients are determined by using data from a 3-month window centered on the target month; e.g., the May analysis consists of April, May, and June (Ivar Seierstad, personal communication, 10 November 2017). The same coefficients, based on mapping the first 24 h, were applied to all lead times and members. For forecasts outside the observation range, a 1 : 1 extrapolation was used. That is, if a forecast is 2 °C higher than the highest mapped forecasted temperature, then the calibrated forecast is 2 °C higher than the highest mapped reference temperature.

For this study, we applied the calibration coefficients provided by MET Norway to the temperature forecasts for the period 2013–2015. Accordingly, ENS was interpolated to the $5$ km $\times 5$ km grid for which the quantile mapping coefficients were used to obtain the calibrated temperature ensembles ($T_{cal}$). Subsequently, the calibrated ensembles on the $5$ km $\times 5$ km grid were downscaled to the $1$ km $\times 1$ km 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 (horizontal). A bias in the ensemble forecast is recognized as a slope in the rank histogram, where a negative slope indicates too-warm temperature forecasts and a positive slope too-cold forecasts. A U shape indicates that the ensemble forecast is under-dispersed, whereas a convex shape indicates over-dispersion (Hamill, 2001). In order to quantify the reliability, a decomposition of the chi-squared 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; i.e., lower values are better, with an optimal value of zero for unbiased 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 temperature sharpness was assessed by first estimating the range between the 5th and the 95th percentile of the ordered ensemble forecasts for all issue dates, lead times, and catchments. For streamflow, we estimated a relative sharpness by dividing the 5th to 95th percentile range by the ensemble mean. Thereafter, sharpness was determined for each catchment and lead time as the average range of all issue dates. The continuous rank probability score (CRPS) is a summary of reliability, sharpness, and uncertainty (Hersbach, 2000). CRPS (denoted as $S_{CRP}$ in Eq. 1) measures the distance between the observation $x_a$ and the ensemble forecast, where the latter is expressed by the cumulative density function $F_x(x)$:
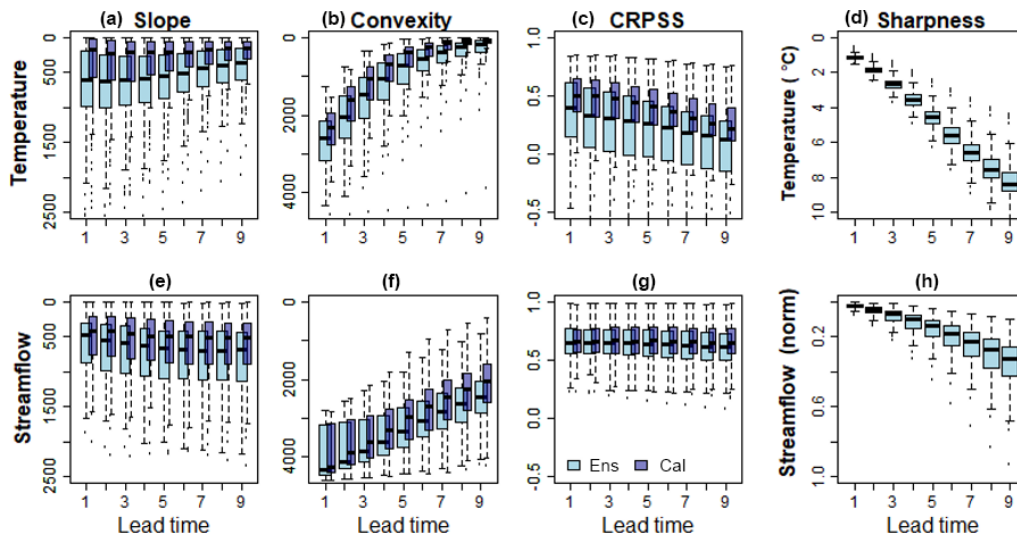
**Figure 3.** Summary of temperature and streamflow scores for all lead times. Each box represents the 139 catchments values averaged over all issue dates. Rank-histogram test decomposition for slope and convexity is shown in **(a)**, **(e)**, **(b)**, and **(f)**, respectively, and CRPSS in **(c)** and **(g)**. Panels **(d)** and **(h)** show sharpness for the uncalibrated forecasts. Temperature is shown in panels **(a)**–**(d)** and streamflow in **(e)**–**(h)**. Results are based on the full dataset and are shown for both uncalibrated (light blue) and calibrated (blue) ensembles at lead times of 1 to 9 days. For slope and convexity, 0 is the optimal value, and the scales are reversed so that the optimal value is on the top, corresponding to the CRPSS optimal value at 1.0.

$$S_{\text{CRP}}(F_x, x_a) = \int_{-\infty}^{\infty} [F_x(x) - H(x - x_a)]^2 \mathrm{d}x, \qquad (1)$$

where $H$ is the Heaviside function that is zero when the argument is less than zero, and one otherwise (Hersbach, 2000). $\overline{\text{CRPS}}$ was calculated as the average CRPS ($S_{\text{CRP}}$) over the study period (1 March 2013 to 31 January 2015). $\overline{\text{CRPS}}$ is similar to the mean absolute error for deterministic forecasts. The temperature $\overline{\text{CRPS}}$ was computed using the seNorge temperature $T_o$ as observations, whereas streamflow $\overline{\text{CRPS}}$ used $Q_{o[c,t]}$ as observations. This evaluation approach allowed us to evaluate the isolated 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 denoted as $S_{\text{CRPS}}$ in Eq. 2), a benchmark is needed. The benchmark is a reference forecast which a skillful forecast should outperform. For both temperature and streamflow, ensembles representing daily climatology were used as benchmarks. Daily seNorge temperatures ($T_{o[c,t]}$) 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 reference 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 ($S_{\text{CRPS}}$) was calculated for each catchment according to Eq. (2) (Hersbach, 2000). CRPSS ($S_{\text{CRPS}}$) was calculated for each catchment according to Eq. (2) (Hersbach, 2000), where $\overline{S}_{\text{B\_CRP}}$ is the benchmark score and $\overline{S}_{\text{F\_CRP}}$ is the forecast score (denoted as $\overline{\text{CRPS}}$ in the text, calculated for the benchmark and forecast, respectively).

$$S_{\text{CRPS}} = \frac{\overline{S}_{\text{B\_CRP}} - \overline{S}_{\text{F\_CRP}}}{\overline{S}_{\text{B\_CRP}}} \qquad (2)$$

CRPSS varies from $-\infty$ to 1, where 1 is a perfect score. Negative values mean that the forecast performs worse than climatology, and CRPSS equal to 0 implies that it performs similarly to the benchmark (climatology in this case). 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 ($T_{\text{ens}}$) and the calibrated ($T_{\text{cal}}$) ensembles to generate the streamflow ensembles. For readability, the abbreviations $S_{\text{CRP}}$ and $S_{\text{CRPS}}$ used in the equation will be substituted with CRPS and CRPSS in the text hereafter.

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 used linear regression to identify relationships between

catchment characteristics (elevation difference and catchment area) and the skill score ($T_{cal}$ and $Q_{cal}$ CRPSS). The linear regression analysis was done for combinations of seasons and regions. Seasonal variations in skill score were assessed by calculating CRPSS for the two seasons, spring (April to June) and autumn (October to December). This definition of seasons is used to better capture a snowmelt season, which for most Norwegian catchments is in the period April to June. For this paper, we chose to focus on the results for autumn and spring. Summer (July to September) was excluded due to the relatively small changes in CRPSS explained by the following: (i) the skill of uncalibrated temperature forecasts is higher and the potential for improvement is lower; and (ii) there is less or no snow in summer, resulting in a reduced streamflow sensitivity to temperature. Winter (January to March) was excluded since it performs similarly to autumn.

Finally, the effect 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 flood thresholds are catchment specific and calculated return periods are based on reference streamflow, which is also the approach used herein.

# 4 Results

Temperature and streamflow forecasts were estimated for 139 catchments, 1036 issue dates, and nine lead times. Figure 3 presents a summary of the validation scores, CRPSS, and the rank-histogram decomposition, in addition to sharpness, for all lead times. Each box plot shows the variations in the validation scores between the catchments. The rank-histogram slope and convexity describe bias and dispersion in the forecasts, respectively; both can be considered a measure for the reliability. As shown in Fig. 3, temperature slope and convexity improve with increasing lead time, whereas CRPSS and sharpness get poorer. For streamflow, slope and sharpness get poorer and 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 on reliability, uncertainty, and sharpness and enables a comparison between catchments. A lead time of 5 days was chosen since reliability (convexity and slope) has improved and some sharpness is maintained; i.e., a too-large ensemble spread will increase the reliability but the forecast value will be reduced.

## 4.1 Temperature forecasts

Time series of seNorge daily temperature $T_o$, the range of raw (uncalibrated) temperature ensembles $T_{ens}$ (left panels), and scatter plots of ensemble mean for both raw $T_{ens}$ and

calibrated $T_{cal}$ versus $T_o$ (right panels) are shown for three selected catchments in Fig. 4. For Gjuvaa (E), a high-altitude catchment (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 raw temperature forecasts for this catchment are representative for most catchments in eastern Norway. Representing western Norway, raw $T_{ens}$ in Viksvatn (W) has a seasonal cold bias that is reduced by the temperature calibration. The cold bias is typical for several catchments in the coastal west, middle, and north regions. Another western catchment, Foennerdalsvatn (W), has a similar cold bias in $T_{ens}$ to Viksvatn (W), but for Foennerdalsvatn the bias is notable for all seasons and even increases for $T_{cal}$ (Fig. 4).

## 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 ($T_{cal}$ and $T_{ens}$) and streamflow ($Q_{cal}$ and $Q_{ens}$) ensembles are shown in Fig. 5. Each dot represents a catchment and the color indicates the region. The two panels in Fig. 5 show how the change in temperature CRPSS affects the change in streamflow CRPSS for spring and autumn. For spring, the relationship is close to the 1 : 1 line, whereas for autumn streamflow is less sensitive to the temperature calibration.

Catchment CRPSSs 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 perform 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 is 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. Figures 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).
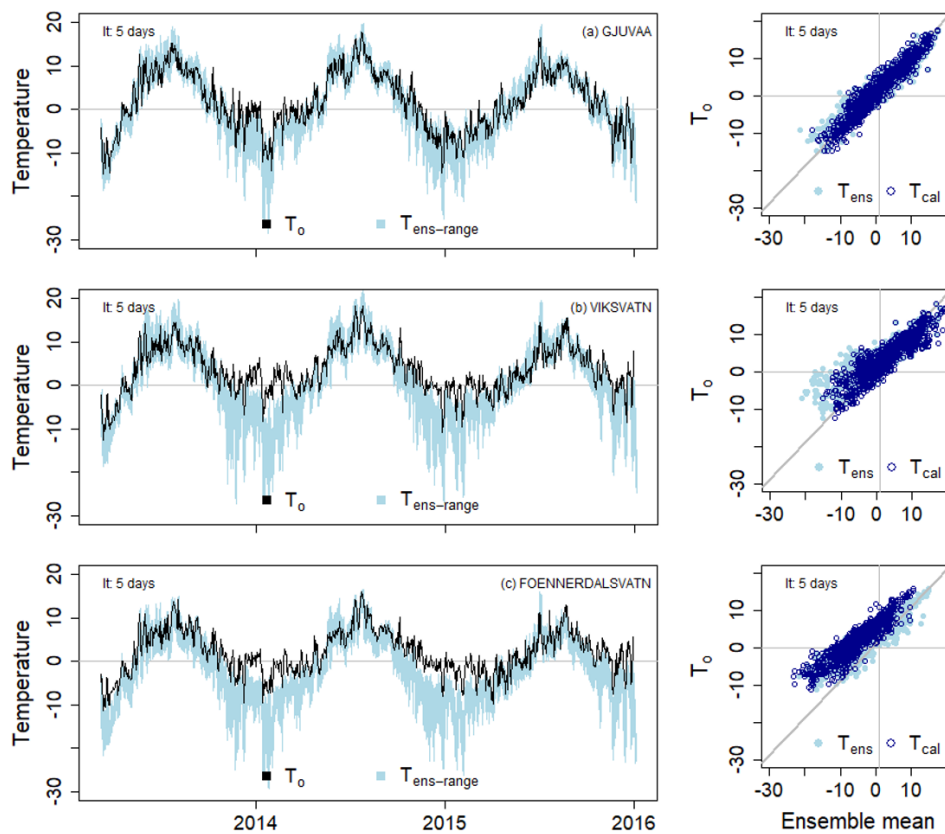
**Figure 4.** Time series of temperature for Gjuvaa **(a)**, Viksvatn **(b)**, and Foennerdalsvatn **(c)** showing the range of uncalibrated temperature ensemble forecast ($T_{\text{ens-range}}$, light blue area) for the period 2013–2015; seNorge observations are shown as black lines. Scatter plots show ensemble mean temperature for both calibrated ($T_{\text{cal}}$, blue) and uncalibrated ($T_{\text{ens}}$, light blue) temperature plotted against seNorge temperature ($T_{\text{o}}$). Lead time (lt) is 5 days.
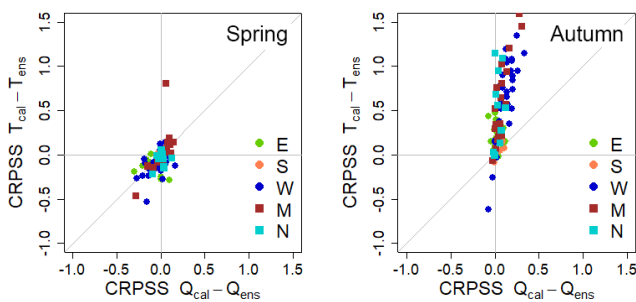


**Figure 5.** Difference in CRPSS for uncalibrated and calibrated temperature for spring and autumn. The difference in temperature skill is plotted on the $y$ axis and the difference in streamflow skill on the $x$ axis. The grey diagonal represent the 1 : 1 line. Catchment values are color indexed by region. All plots are presented for a lead time of 5 days.

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 (5 % sig-

nificance level). Elevation difference is negatively correlated to streamflow CRPSS for the east and middle regions. The east region also has a negative correlation between streamflow CRPSS and catchment area as opposed to the other regions that have a positive correlation. The correlation does not change sign between the seasons for any of the regions. Calibrated temperature and streamflow CRPSS plotted as a function of catchment area are presented for the east and south in Fig. 9.

### 4.3 Snowmelt flood 2013

Forecasts and observations for a snowmelt-driven flood are presented in Fig. 10 for Bulken (W), located in western Norway. The figure shows forecasted streamflow for lead times of 2, 5, and 9 days for the target dates 16–26 May 2013. Note that for the lead times of 2, 5, and 9 days, the forecasts for 18 May are issued on 16, 13, and 9 May, respectively. The horizontal grey dotted lines represent the mean annual, the 5-year, and the 50-year floods (i.e., the operational flood warning levels) in this catchment. 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
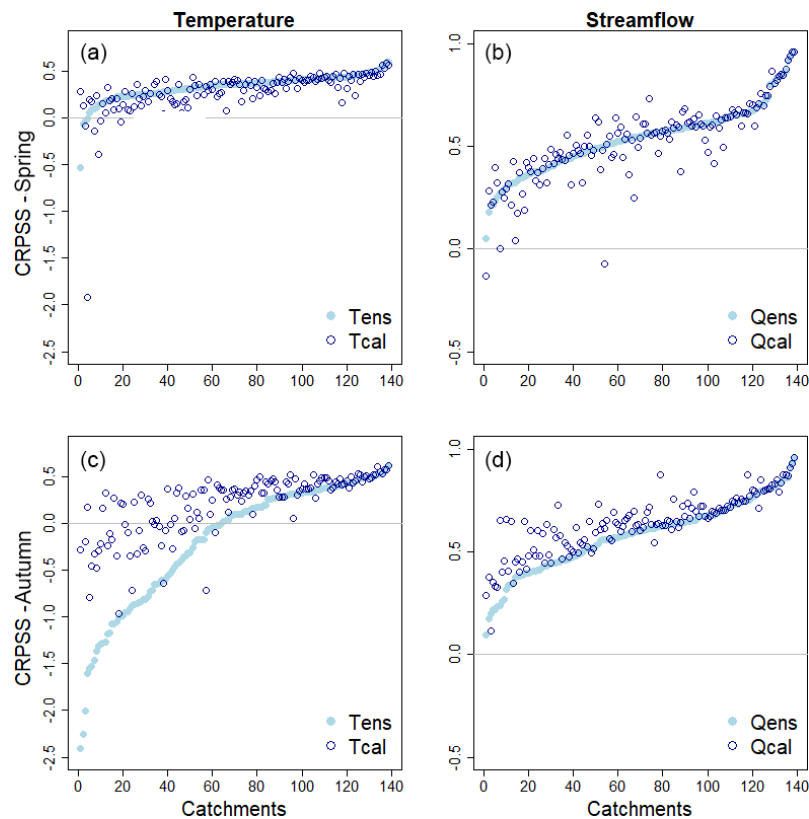
**Figure 6.** Temperature ($T_{ens}$ and $T_{cal}$, **a, c**) and streamflow ($Q_{ens}$ and $Q_{cal}$, **b, d**) CRPSS for SPRING (**a, b**) and AUTUMN (**c, d**). The catchments are ordered by increasing CRPSS for $T_{ens}$ and $Q_{ens}$ (light blue dots); the catchment calibrated values ($T_{cal}$ and $Q_{cal}$) are plotted as blue circles. All results are presented for a lead time of 5 days.

**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 ($\Delta H$), for the five regions. "✓" indicates a significant positive relationship, "×" a significant negative relationship, and "ns" a non-significant relationship. Results are for a lead time of 5 days.

|  |  | $T_{cal}$ | $Q_{cal}$ | $T_{cal}$ | $Q_{cal}$ |
|---|---|---|---|---|---|
|  |  | SPRING | | AUTUMN | |
| Area (km$^2$) | East | ns | × | ns | × |
|  | South | ✓ | ✓ | ns | ✓ |
|  | West | ✓ | ns | ns | ns |
|  | Middle | ✓ | ns | ✓ | ns |
|  | North | ns | ns | ✓ | ns |
| $\Delta H$ (m) | East | ns | × | ns | × |
|  | South | ns | ns | ns | ns |
|  | West | ns | ns | ns | ns |
|  | Middle | ns | ns | ns | × |
|  | North | ns | ns | ns | ns |

see how the ensemble spread increases with lead time (from lower to upper panel), from a narrow range around the ensemble mean for the lead time of 2 days to a very wide range for a lead time of 9 days.

## 5  Discussion

Box plots of validation scores for all catchments and lead times in Fig. 3 show that, on average, both raw $T_{ens}$ and calibrated $T_{cal}$ temperature ensembles 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) as measured by rank-histogram slope and convexity improved with longer lead times, the reduced sharpness and increased uncertainty resulted in a reduced skill (CRPSS). For streamflow, the bias increased with longer lead times, 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 most validation scores and lead times. The calibration procedure applied in this study involves many interpolations and downscaling steps that increase the uncertainty in temperature forecasts. We be-
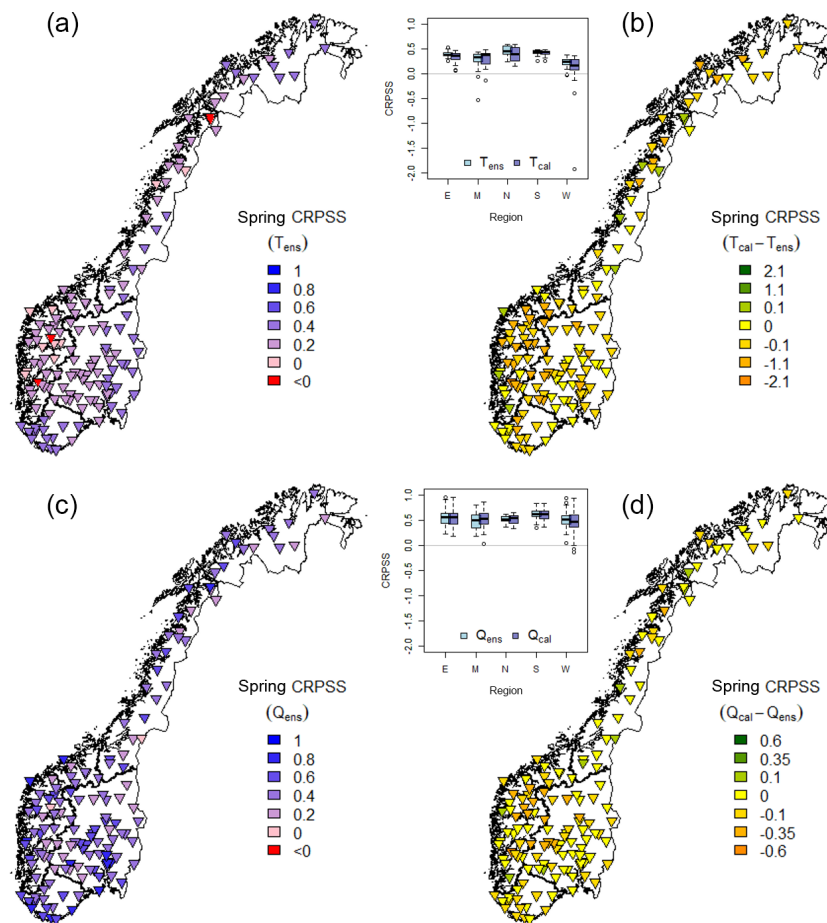
**Figure 7.** Spring CRPSS for uncalibrated forecasts **(a, c)** and CRPSS difference between calibrated and uncalibrated forecasts **(b, d)** for temperature **(a, b)** and streamflow **(c, d)**. A darker blue color **(a, c)** indicates an optimal performance (maximum CRPSS = 1.0), pink a CRPSS of zero, and red a negative value. A green color **(b, d)** indicates a positive effect of temperature calibration on the skill, yellow means no effect, and an orange color indicates a negative effect. The box plots show temperature and streamflow CRPSSs grouped by region (Fig. 1). All results are presented for a lead time of 5 days.

lieve that a catchment-specific temperature calibration, tailored to the needs of hydrological forecasting, would solve this challenge.

## 5.1 Effect of temperature calibration for the temperature forecast skill

The skill for both raw (uncalibrated) $T_{ens}$ and calibrated $T_{cal}$ temperature ensembles varies with season (Figs. 5–8). The relatively small temperature skill improvements in spring and large skill improvements in autumn 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 a cold bias and lays the foundation 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, as most statistical techniques, is sensitive to forecasts outside the range of calibration values and periods (Lafon et al., 2013), which can be an explanation for too high a correction in the highest $T_{ens}$ quantile. The use of forecasts from different model cycles might affect the consistency in the forecasts. Moreover, the calibration parameters are sensitive to the representativeness of the calibration period.

The most pronounced spatial pattern is the low autumn CRPSS for uncalibrated ensembles $T_{ens}$ in the coastal areas. This is seen in the box plots for the west, middle, and north regions (Fig. 8) and in the plots of the western catchments Viksvatn and Foennerdalsvatn during winter months (Fig. 4). This cold bias is documented for the Norwegian coastal areas in the cold seasons by Seierstad et al. (2016) and is mainly caused by the radiation calculations in the ECMWF model
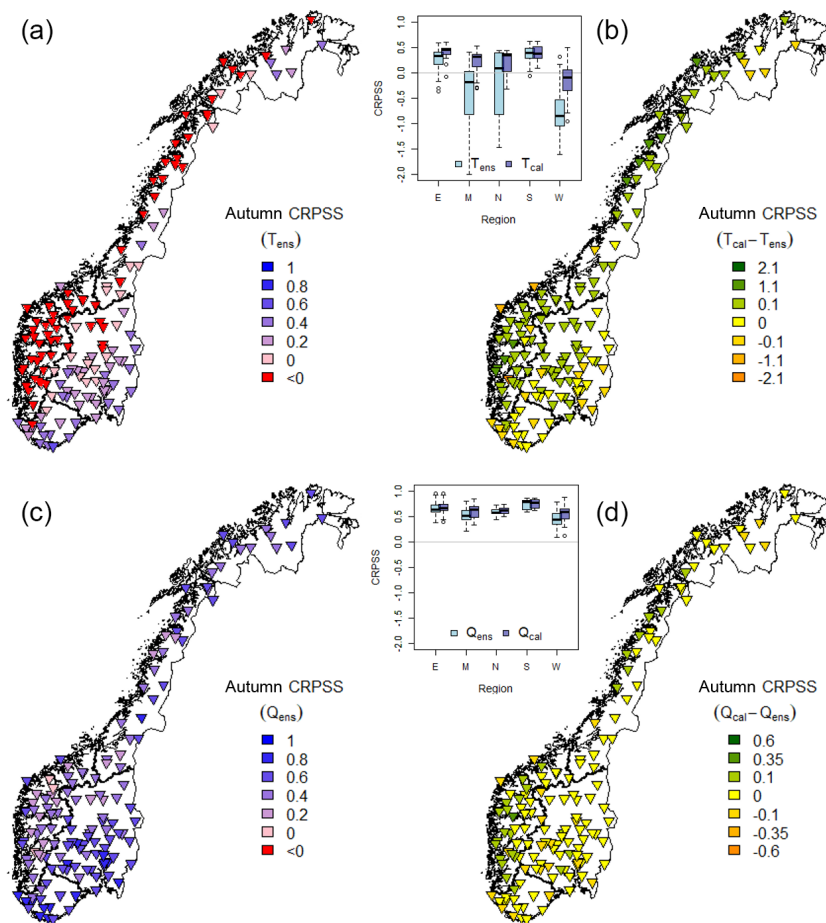
**Figure 8.** Autumn CRPSSs for uncalibrated forecasts are presented in panels **(a)** and **(c)**, where the darker blue color indicates an optimal performance (maximum CRPSS = 1.0), pink color represents a CRPSS of zero, and red a negative value. The differences in CRPSS between calibrated and uncalibrated forecasts are presented in panels **(b)** and **(d)**, where the green color indicates a positive effect of temperature calibration on the skill, yellow zero, and orange color indicates a negative effect. Temperature CRPSS is presented in **(a)** and **(b)** and streamflow CRPSS in **(c)** and **(d)**. The box plots of both calibrated and uncalibrated temperature and streamflow CRPSS show catchments grouped by region (Fig. 1). All results are presented for a lead time of 5 days.

(Hogan et al., 2017). The coarse radiation grid results in warmer sea points being used to compute longwave fluxes applied over colder land points, causing too much cooling. This effect is seen for the temperature forecasts for winter 2014 and 2015 for the coastal catchments in Fig. 4b and c, in contrast to the inland catchment (Fig. 4a) which is less biased. The radiation resolution is improved in later model cycles (Hogan et al., 2017; Seierstad et al., 2016). 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 south and east regions, CRPSS shows that the uncalibrated $T_{ens}$ is skillful for both autumn and spring; hence, the calibration has a smaller effect in these catchments.

## 5.2  Effect of temperature calibration for the streamflow forecast skill

The skill of the temperature-calibrated streamflow ensemble forecasts, $Q_{cal}$, improved for most of the catchments for autumn, while both improved and reduced skill were seen for spring (Figs. 5–8). Autumn streamflow skill was improved by temperature calibration for all regions, the largest improvement was seen for the coast and the west and middle regions. Two possible explanations for this spatial pattern are (i) the improvement in temperature forecast skill during autumn in these regions and (ii) that many coastal catchments are more sensitive to the calibration of temperatures since the temperatures are more frequently around 0 °C compared to the colder and dryer inland catchments. In spring, no clear spatial patterns are seen, neither for $Q_{ens}$ nor for the change in skill.

It is also evident that, independent of the sign of the temperature skill change (Fig. 5), a change in temperature has a
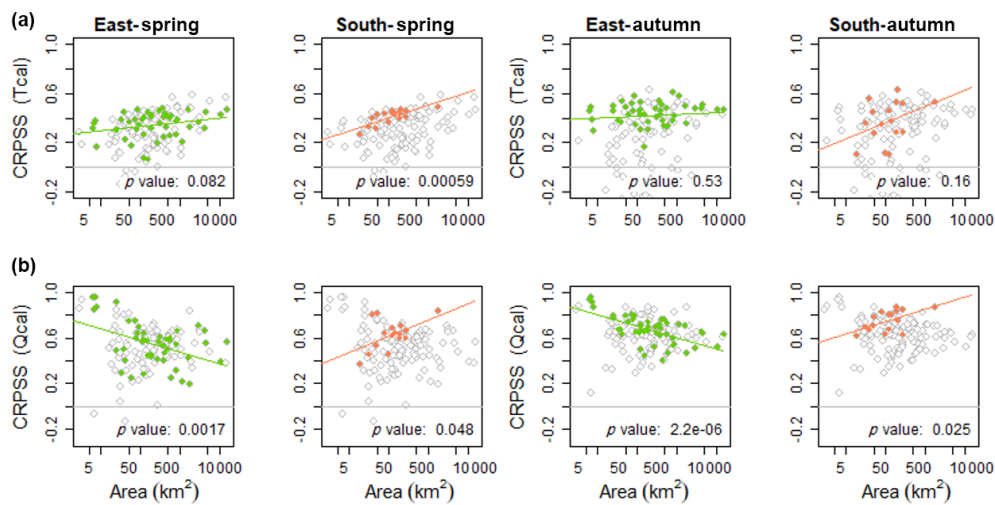
**Figure 9.** Temperature (row **a**) and streamflow (row **b**) CRPSS for the east (E) and south (S) regions, plotted as a function of catchment area for both autumn and spring. The colored dots show the CRPSS for the respective regions whereas the grey circles show the CRPSS for all 139 catchments. 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 a lead time of 5 days.

larger impact on streamflow in spring than in autumn. During spring, temperatures are often close to the two threshold temperatures that control the phase of precipitation and the onset of snowmelt. Such periods are challenging to simulate correctly (Engeland et al., 2010). Of additional importance, for spring as opposed to autumn, is the snow storage at the end of winter, as well as the snowmelt contribution to streamflow. Hence, estimated streamflow has 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 preprocessing temperature and precipitation for the streamflow skill in the Rhine catchments. The results presented herein and in the cited papers indicate that the effect of preprocessing depends on the hydrological regime (i.e., sensitivity to temperature), the initial skill of the forcing variables, and on which temporal periods (i.e., for specific events, seasons, or the whole year) the sensitivity is evaluated. The same lead time was used to relate improvement in streamflow to temperature; we consider this robust since most catchments in this study have a concentration time of less than a day.

In summary, it can be concluded that, to further improve streamflow forecasts during the snowmelt season, improved temperature forecasts are essential. Streamflow forecasts during spring have the highest potential for improvement since the temperature forecasts were 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 forecasts, but the sensitivity is less than for spring.

## 5.3 Catchment characteristics and skill

Only a few significant relationships between the catchment characteristics, e.g., catchment area and elevation gradient, 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 were found for any combination of region or season. A positive relationship between temperature skill and catchment area was found for 5 out of 10 regression equations. This result is not conclusive, but indicates that (i) the smallest catchments are smaller than the grid size of the ECMWF model and therefore sensitive to the preprocessing and (ii) it is more challenging to forecast weather on small spatial scales than large spatial scales.

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 3 out of 10 datasets. This suggests that the downscaling approach has the 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.

Forecasting in small catchments with particular characteristics may be challenging since they may not be well represented, neither by the numerical weather prediction model nor by the calibration methods. In our dataset, Foennerdalsvatn (Fig. 4c) is such an example. The catchment area is only
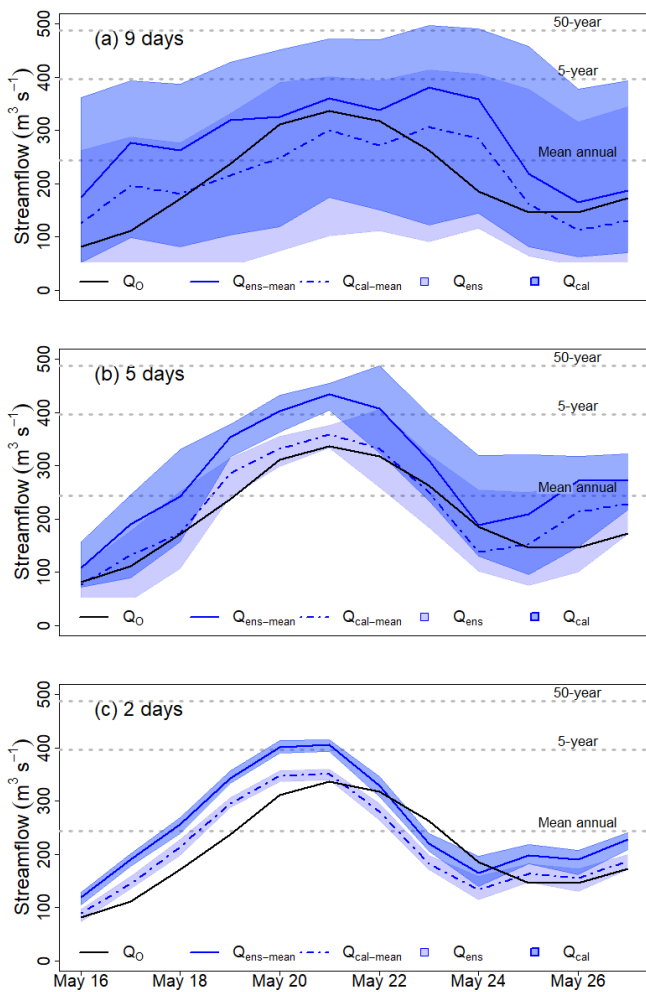
**Figure 10.** Forecasted streamflow for the Bulken catchment fort lead times of 9, 5, and 2 days. Forecast target dates on the $x$ axis, and streamflow ($m^3 s^{-1}$) on the $y$ axis. Reference streamflow with seNorge observations $Q_o$ (black solid line), ensemble mean uncalibrated temperature $Q_{ens}$ (blue line), ensemble mean calibrated $Q_{cal}$ (blue dotted line), ensemble range $Q_{ens}$ (light violet area), and ensemble range $Q_{cal}$ (light blue area). The grey dotted lines indicate the thresholds for mean annual, 5-year, and 50-year floods.

$7.1 \, km^2$, elevation is high, topography is steep, glaciers cover 47 % of the catchment area, and it is located close to the coast.

### 5.4   Snowmelt flood 2013

The snowmelt flood event (Fig. 10) illustrates clearly how temperature calibration affects forecasted ensemble 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 below to above the 5-year flood. For this event, however, the use of calibrated temperature reduced the performance of the forecasted streamflow, $Q_{cal}$. The refer-

ence streamflow, $Q_o$, is better captured by the streamflow forecasts based on uncalibrated temperature forecasts, $Q_{ens}$. The deterioration in the forecast performance using calibrated temperature is particular for this event. Other results provided in this study show clearly that the calibrated temperature ensembles improve the streamflow forecasts on average.

Figure 10 reveals how the ensemble range for the snowmelt event clearly increases with increasing lead time. For a lead time of 2 days (lower panel), the range is too narrow, while for a lead time of 9 days (upper 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, as well as to identify potential improvements in the forecasting chain. We applied a gridded temperature calibration method and evaluated its effect 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 poorly performing coastal catchments in autumn and winter.

- The effect of calibration on temperature skill was less clear in spring. In spring, the calibrated temperature resulted in reduced skill for many catchments.

- Smaller bias in spring 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 effect on streamflow skill, compared to autumn and winter.

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 catchments along the coast in the west, middle,

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.

Is streamflow affected by temperature calibration during a 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 dependent 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 preprocessing of temperature includes both an elevation correction depending 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. In this study, we have investigated the isolated effect of uncertainties in temperature forecasts. For a more complete assessment of forecast uncertainties, error in initial conditions, hydrological model parameters, and structure need to be accounted for. In particular, we might expect a strong interaction between uncertainties in temperature forecasts and model parameters controlling snow accumulation and snowmelt processes.

The conclusions in this study are based on a testing period of almost 3 years. Even if this is a relatively short testing period, 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. 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 snowmelt season. In addition, this study shows that reducing the cold temperature bias in coastal areas results in improved streamflow forecasts and that the preprocessing needs to account for seasonal differences in temperature forecasts (biases).

## References

Aguado, E. and Burt, J. E.: Understanding weather and climate, 5th Edn., Upper Saddle River, NJ, USA, Pearson Prentice Hall, 2010.

Beldring, S.: Distributed Element Water Balance Model System, report 4, 40 pp., Norwegian Water Resources and Energy directorate, 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, Mon. Weather Rev., 145, 1919–1935, https://doi.org/10.1175/mwr-d-16-0417.1, 2017.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, Swedish Meteorological and Hydrological Institute SMHI, Report No. RHO 7, Norrköping, Sweden, 1976.

Bremnes, J. B.: Improved calibration of precipitation forecasts using ensemble techniques. Part 2: Statistical calibration methods, met.no, Report no. 4, 34 pp., Oslo, Norway, available at: http://met-xpprod.customer.enonic.io/publikasjoner/met-report/met-report-2007 (last access: 1 February 2019), 2007.

Buizza, R.: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system, Mon. Weather Rev., 125, 99–119, https://doi.org/10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2, 1997.

Buizza, R., Milleer, M., and Palmer, T. N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, Q. J. Roy. Meteor. Soc., 125, 2887–2908, https://doi.org/10.1002/qj.49712556006, 1999.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M.: Comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems, Mon. Weather Rev., 133, 1076–1097, 2005.

Cloke, H. L. and Pappenberger, F.: Ensemble Forecasting: A review, J. Hydrol., 375, 613–626, 2009.

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, Nat. Hazards Earth Syst. Sci., 13, 1051–1062, https://doi.org/10.5194/nhess-13-1051-2013, 2013.

Dobrovičová, S., Dobrovič, R., and Dobrovič, J.: The Economic Impact of Floods and their Importance in Different Regions of the World with Emphasis on Europe, Proc. Econ. Financ., 34, 649–655, https://doi.org/10.1016/S2212-5671(15)01681-0, 2015.

Doherty, J.: Calibration and Uncertainty Analysis for Complex Environmental Models, Watermark Numerical Computing, Brisbane, Australia, ISBN 978-0-9943786-0-6, 2015.

ECMWF: Set III – Atmospheric model Ensemble 15-day forecast (ENS), available at: https://www.ecmwf.int/en/forecasts/datasets/set-iii (last access: 20 December 2018), 2018a.

ECMWF: Changes in ECMWF models, available at: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model (last access: 20 December 2018), 2018b.

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

Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the HBV model, J. Hydrol., 384, 142–155, 2010.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynam., 53, 343–367, 2003.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon. Weather Rev., 133, 1098–1118, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. R. Stat. Soc. B, 69, 243–268, https://doi.org/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, Hydrol. Res., 10, 273–286, 1979.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001.

Hamill, T. M. and Colucci, S. J.: Verification of Eta-RSM Short-Range Ensemble Forecasts, Mon. Weather Rev., 125, 1312–1327, 1997.

Hamill, T. M.: Comments on "Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging", Mon. Weather Rev., 135, 4226–4230, https://doi.org/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, Oslo, Norway, 2017.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, https://doi.org/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, available at: https://www.ecmwf.int/en/elibrary/17771-radiation-numerical-weather-prediction (last access: 20 December 2018), 2017.

Jolliffe, I. T. and Primo, C.: Evaluating rank histograms using decompositions of the chi-square test statistic, Mon. Weather Rev., 136, 2133–2139, 2008.

Lafon, T., Dadson, S., Buys, G., and Prudhomme, C.: Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods, Int. J. Climatol., 33, 1367–1381, https://doi.org/10.1002/joc.3518, 2013.

Langsrud, Ø., Frigessi, A., and Høst, G.: Pure model error for the HBV model, NVE, Hydra notat 4/98, 28 pp., Oslo, Norway, 1998.

Langsrud, Ø., Høst, G., Follestad, T., Frigessi, A., and Hirst, D.: Quantifying uncertainty in HBV runoff forecasts by stochastic simulations, NVE, Hydra notat 2/99, 38 pp., Oslo, Norway, 1999.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, J. Comput. Phys., 227, 3515–3539, 2008.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z.: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting, WIRES Water, 4, e1246, https://doi.org/10.1002/wat2.1246, 2017.

Mohr, M.: New routines for gridding of temperature and precipitation observations for "seNorge.no", met.no, Note no. 8, 40 pp., Oslo, Norway, 2008.

Müller, M., Homleid, M., Ivarsson, K.-I., Køltzow, M. A. Ø., Lindskog, M., Midtbø, K. H., Andrae, U., Aspelien, T., Berggren, L., Bjørge, D., Dahlgren, P., Kristiansen, J., Randriamampianina, R., Ridal, M., and Vignes, O.: AROME-MetCoOp: A Nordic Convective-Scale Operational Weather Prediction Model, Weather Forecast., 32, 609–627, https://doi.org/10.1175/waf-d-16-0099.1, 2017.

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

Owens, R. G. and Hewson, T. D.: ECMWF Forecast User Guide, ECMWF, Reading, UK, https://doi.org/10.21957/m1cs7h, 2018.

Pagès, M. and Miró, J. R.: Determining temperature lapse rates over mountain slopes using vertically weighted regression: a case study from the Pyrenees, Meteorol. Appl., 17, 53–63, 2010.

Peel, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11, 1633–1644, https://doi.org/10.5194/hess-11-1633-2007, 2007.

Persson, A.: User guide to ECMWF forecast products, edited by: Andersson, E. and Tsonevsky, I., ECMWF, Reading, UK, 2015.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, 2005.

Rizzi, J., Nilsen, I. B., Stagge, J. H., Gisnås, K., and Tallaksen, L. M.: Five decades of warming: impacts on snow in Norway, Hydrol. Res., 49, 670–688, https://doi.org/10.2166/nh.2017.051, 2017.

Schaake, J., Pailleux, J., Thielen, J., Arritt, R., Hamill, T., Luo, L., Martin, E., McCollor, D., and Pappenberger, F.: Summary of recommendations of the first workshop on Postprocessing and Downscaling Atmospheric Forecasts for Hydrologic Applications held at Météo-France, Toulouse, France, 15–18 June 2009, Atmos. Sci. Lett., 11, 59–63, 2010.

Seierstad, I., Kristiansen, J., and Nipen, T.: Better temperature forecasts along the Norwegain coast, Newsletter, 148, available at: https://www.ecmwf.int/en/newsletter/148/news/better-temperature-forecasts-along-norwegian-coast (last access: 1 February 2019), 2016.

Sheridan, P., Smith, S., Brown, A., and Vosper, S.: A simple height-based correction for temperature downscaling in complex terrain, Meteorol. Appl., 17, 329–339, 2010.

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, 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, Meteorol. Appl., 12, 57–68, https://doi.org/10.1017/S1350482705001490, 2005.

Vannitsem, S., Wilks, D. S., and Messner, J. W. (Eds.): Statistical Postprocessing of Ensemble Forecasts, Elsevier, ISBN 9780128123720, https://doi.org/10.1016/B978-0-12-812372-0.09988-X, 2018.

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, J. Hydrol., 501, 73–91, 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, 913–931, https://doi.org/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, J. Hydrol., 538, 33–48, 2016.

Wang, X. and Bishop, C. H.: Improvement of ensemble reliability with the new dressing kernel, Q. J. Roy. Meteor. Soc., 131, 965–986, 2005.

Wetterhall, F., Pappenberger, F., Alfieri, L., Cloke, H. L., Thielen-del Pozo, J., Balabanova, S., Danhelka, J., Vogelbacher, A., Salamon, P., Carrasco, I., Cabrera-Tordera, A. J., Corzo-Toscano, M., Garcia-Padilla, M., Garcia-Sanchez, R. J., Ardilouze, C., Jurela, S., Terek, B., Csik, A., Casey, J., Stankunavicius, G., Ceres, V., Sprokkereef, E., Stam, J., Anghel, E., Vladikovic, D., Alionte Eklund, C., Hjerdt, N., Djerv, H., Holmberg, F., Nilsson, J., Nyström, K., Sušnik, M., Hazlinger, M., and Holubecka, M.: HESS Opinions "Forecaster priorities for improving probabilistic flood forecasts", Hydrol. Earth Syst. Sci., 17, 4389–4399, https://doi.org/10.5194/hess-17-4389-2013, 2013.

Wilks, D. S. and Hamill, T. M.: Comparison of Ensemble-MOS Methods Using GFS Reforecasts, Mon. Weather Rev., 135, 2379–2390, 2007.

Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R.: Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging, Mon. Weather Rev., 135, 1364–1385, https://doi.org/10.1175/MWR3347.1, 2007.

Paper II:

# The benefits of pre- and postprocessing streamflow forecasts for an operational flood-forecasting system of 119 Norwegian catchments

**T. J. Hegdahl, K. Engeland, A. Singleton, and I. Steinsland**

II

1 **The benefits of pre- and postprocessing streamflow forecasts for**

2 **an operational flood-forecasting system of 119 Norwegian**

3 **catchments**

4

5 Trine J. Hegdahl[1], Kolbjørn Engeland[1,2], Ingelin Steinsland[3], Andrew Singleton[4]

6 [1]Norwegian Water Resources and Energy Directorate, Hydrological Modelling, 0301 Oslo, Norway
7 [2]University of Oslo, Department of Geosciences, 0316 Oslo, Norway
8 [3]Norwegian University of Science and Technology, Department of Mathematical Sciences, 7034 Trondheim, Norway
9 [4]Norwegian Meteorological Institute, 0313 Oslo, Norway
10

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

12

13 **Abstract**. The novelty of this study is to evaluate the univariate and the combined effects of including both

14 precipitation and temperature forecasts in the preprocessing together with the postprocessing of streamflow for

15 forecasting of floods as well as all streamflow values for a large sample of catchments. A hydrometeorological

16 forecasting chain in an operational flood forecasting setting with 119 Norwegian catchments was used. This study

17 evaluates the added value of pre- and postprocessing methods for ensemble forecasts in a hydrometeorological

18 forecasting chain in an operational flood forecasting setting with 119 Norwegian catchments. Two years of ECMWF

19 ensemble forecasts of temperature (T) and precipitation (P) with a lead-time up to 9 days were used to force the

20 operational hydrological HBV model to establish streamflow forecasts. Two approaches to preprocess the temperature

21 and precipitation forecasts were tested. 1) An existing approach applied to the gridded forecasts using quantile mapping

22 for temperature and a Bernoulli-gamma distribution for precipitation. 2) Bayesian model averaging (BMA) applied to

23 catchment average values of temperature and precipitation. BMA was also used for postprocessing catchment

24 streamflow forecasts. Ensemble forecasts of streamflow were generated for a total of fourteen schemes based on

25 combinations of raw, preprocessed, and postprocessed forecasts in the hydrometeorological forecasting chain. The aim

26 of this study is to assess which pre- and postprocessing approaches should be used to improve streamflow and flood

27 forecasts and look for regional or seasonal patterns in preferred approaches.

28

29 The forecasts were evaluated for two datasets: i) all streamflows and ii) flood events with streamflow above mean

30 annual flood. Evaluations were based on reliability, continuous ranked probability score (CRPS) and -skill score

31 (CRPSS). For the flood dataset, the critical success index (CSI) was used. Evaluations based on all streamflow data

32 showed that postprocessing improved the forecasts only up to a lead-time of two to three days, whereas preprocessing

33 T and P using BMA improved the forecasts for 50% - 90% of the catchments beyond three days lead-time. However,

34 for flood events, the added value of pre- and postprocessing is smaller. Preprocessing of P and T gave better CRPS for

35 marginally more catchments compared to the other schemes.

1 Based on CSI, we found that many of the forecast schemes perform equally well. Further, we found large differences
2 in the ability to issue warnings between spring and autumn floods. There was almost no ability to predict autumn floods
3 beyond 3 days, whereas the spring floods had predictability up to 9 days for many events and catchments. The results
4 indicate that the ensemble forecasts have problems in predicting correct autumn precipitation, and the uncertainty is
5 larger for heavy autumn precipitation compared to spring events when temperature driven snow melt is important. To
6 summarize we find that the flood forecasts benefit from most pre- and postprocessing schemes, although the best
7 processing approaches depend on region, catchment, and season, and that the processing scheme should be tailored to
8 each catchment, lead time, season, and the purpose of the forecasting.

9 **1    Introduction**

10 Floods can have severe economic, personal, and social costs. Early warnings based on flood forecasts enable both the
11 management authorities and the public to take necessary measures to reduce the impact of floods (e.g., UNISDRI,
12 2004, Pappenberger et al., 2015). However, predicting the future is adhered with uncertainty. Attaching the forecast
13 uncertainty to a predicted flood level adds value for many end users allowing them to do risk evaluation in light of
14 their often-unique circumstances, and thus take measures that are most appropriate and cost effective for them.

15 In the hydro-meteorological forecasting chain there are multiple sources to uncertainty. There is uncertainty in
16 observations, initial conditions, forcing data, model description, and model parameters (e.g., Buizza et al., 1999; Zappa
17 et al., 2011). For flood forecasting an important source of uncertainty and errors are the forcing in the forecasting
18 period, i.e. precipitation and temperature weather forecasts (e.g. Zappa et al., 2011), and this is the focus of this paper.

19 From weather prediction systems it is known that small changes in the initial conditions will affect atmospheric
20 trajectories and future weather predictions (e.g., Lorenz, 1969; Buizza, 2008). To capture the uncertainty in weather
21 prediction caused by initial conditions and model parametrization, ensemble prediction systems (EPS) were developed
22 as early as the 70s (Leith, 1974). The use of meteorological ensembles as input to hydrological models is one approach
23 to achieve probabilistic streamflow forecasts, and thereby provide a probability of the forecasted flood to exceed a
24 given level (Buizza, 2008).

25 Today, ensemble weather forecasts are available as operational services, and using these for hydrological forecasts
26 have been studied in the literature, see e.g., Cloke and Pappenberger (2009) and Wetterhall et al. (2013). To get
27 unbiased and reliable hydrological forecasts, preprocessing (applied to the meteorological forcing) and/or
28 postprocessing (applied to the hydrological output) techniques are needed. Several processing methods are proposed
29 in literature, see e.g., Vannitsem et al. (2018) for an overview. For a national or regional flood forecasting service, a
30 large number of catchments with different hydrological processes and regimes are considered. Therefore, to assess the
31 added value of pre- and postprocessing, a dataset from a large number of catchments that well represent the variability
32 on hydrological processes is needed to provide robust conclusions. In addition, it is important to assess (and compare)
33 the performance of flood forecast, not all streamflow values, for different pre- and postprocessing schemes. In most
34 papers, ensemble forecasts of all streamflow values for one or a small number of catchments are evaluated. This paper
35 aims to fill two knowledge gaps: 1) To gain understanding of the differences in quality for pre and/or post processing
36 method for a range of catchments, and 2) The assess the quality of pre- and postprocessing for flood forecasts.

1   Reliability and accuracy are key characteristics used to measure the quality of ensemble forecasts. A reliable forecast

2   is statistically calibrated (e.g., for 90% of the forecasts, the observations are within the 90% prediction interval). Raw

3   ensemble forecasts are rarely reliable in this sense. The discrepancy between the weather predictions and point

4   measurements shows that forecast ensembles are often biased and underdispersive (Gneiting et al., 2005). A lack of

5   dispersion in global meteorological ensembles is most evident for the shortest lead times and can be explained by

6   slower growth rates of the perturbations in the ensemble prediction system compared to those of an instable "true"

7   atmosphere (Hamill and Colucci, 1997). To correct for bias and underdispersion in the ensemble system, different

8   statistical postprocessing approaches are applied to achieve calibrated ensembles. Li et al. (2017) and Vannitsem et al.

9   (2018) provide a comprehensive review of processing techniques, both parametric approaches relying on parametric

10  probability distributions, for example Bayesian model averaging (BMA) and non-homogeneous Gaussian regression

11  (NGR), and nonparametric approaches like quantile regression and ensemble error dressing methods. Raferty et al.

12  (2005) introduced BMA to the atmospheric community as a statistical method to achieve calibrated and sharp forecasts,

13  and the method has since been widely used within the community (Fraley et al., 2010). More recently studies that use

14  BMA for postprocessing to improve streamflow forecasts have been carried out. For example, Madadgar et al. (2014)

15  used copula embedded BMA for postprocessing streamflow forecasts and improved the forecasts compared to quantile

16  mapping techniques. Jha et al. (2018) demonstrated the use of BMA to remove bias and reduce errors in the

17  precipitation forecasts responsible for a flood event. NGR accounts for the errors in the mean, but unlike an ordinary

18  regression, the error variance is not assumed to be constant, but rather to vary linearly with respect to the ensemble

19  variance (Wilks and Hamill, 2007; Gneiting et al., 2005). Quantile regression applied to ensemble forecasts was

20  introduced by Bremnes (2004) and was first used to correct precipitation forecasts. The method can be viewed as a

21  non-parametric counterpart to NGR, where the predictive probability distribution is described by a set of quantiles.

22  Linear regression is used to describe the relationship between the observations and the forecasts, and the regression

23  parameters are specific for each quantile. There are variations of most methods, and ensemble dressing is one that has

24  both parametric and non-parametric approaches. Roulston and Smith (2003) suggested a non-parametric kernel

25  dressing method, where the kernel represents a distribution of errors from previous forecasts, which is applied to each

26  member of the ensemble. Wang and Bishop (2005) extended this idea and suggested the use of a parametric dressing

27  method of Gaussian kernels where the parameters were estimated by the training data.

28  Previous studies have analyzed the effects of both pre- and postprocessing on short- to medium-range ensemble

29  streamflow forecasts (e.g., Zalachori et al., 2012; Roulin and Vannitsem, 2015; Benninga et al., 2017, Sharma et al.,

30  2018). Few studies include preprocessing of temperature. Verkade et al. (2013), Benninga et al. (2017), and Hegdahl

31  et al. (2019) all applied variations of quantile mapping techniques to calibrate the temperature forecasts, whereas

32  Zalachori et al. (2012) applied an analog approach. Hegdahl et al. (2019) showed that in catchments with seasonal

33  snow cover, temperature calibration is important for improved streamflow forecasts Variations of logistic regression

34  approaches are most common in the studies that preprocessed precipitation (Verkad et al., 2013; Roulin and Vannitsem

35  et al., 2015; Benninga et al., 2017; Sharma et al., 2018). One exception is the analog approach applied by Zalachori et

36  al. (2012). A larger variety of approaches are used to postprocess streamflow; Bayesian processing (Reggiani et al.,

37  2009), Bayesian model averaging including multi-model approaches (Rings et al. 2012; Parish et al. 2012; Xu et al.,

38  2019), variations of quantile regression (Bogner et al., 2016; Benninga et al., 2017; Sharma et al., 2018), extended

39  logistic regression (Fundel and Zappa 2011), and ensemble model output statistics (Roulin and Vannitsem 2015). Some

1    key findings are that calibrated precipitation forecasts do not necessarily lead to calibrated streamflow forecasts

2    (Zalachori et al., 2012; Verkade et al., 2013; Benninga et al., 2017). Postprocessing alone is the simplest way to

3    improve forecasting performance (Zalachori et al., 2012; Sharma, 2018), but not always with a significant improvement

4    (Benninga et al., 2017). Preprocessing the meteorological forcing is important for forecasting high streamflow since

5    errors from the meteorological model are dominant in this case (Benninga et al., 2017). Preprocessing has the highest

6    skill improvement in the warm season, whereas postprocessing is the most effective in the cold season with snow cover

7    (Sharma et al., 2018). This summary indicates that the relative importance of pre- and postprocessing depends on

8    factors including lead time, streamflow magnitude and season.

9    From the literature on short- to medium range streamflow forecasts we have identified two studies investigating the

10    combined effect of preprocessing temperature and precipitation as well as postprocessing the streamflow (Benninga et

11    al., 2017; Zalachori et al., 2012). However, neither of these two studies consider the impacts of such pre- and

12    postprocessing strategies on the forecasting of flood events directly. Zalachori et al. (2012) assess the performance for

13    all streamflows and Benninga et al. (2017) assess the performance for, not necessarily flood inducing, high flows. In

14    Benninga et al. (2017) the forecasts are evaluated for only one catchment, and the author acknowledge that more

15    catchments are needed to verify the generality of their results. A 'large catchment sample' is needed to draw robust

16    conclusions in such studies (Gupta et al., 2014).

17    The two unique contributions of our study are to (i) evaluate the univariate and the combined effects of including both

18    precipitation and temperature forecasts in the preprocessing together with the postprocessing of streamflow for

19    forecasting of both floods as well as all streamflow values, and to (ii) perform the evaluation for a large catchment

20    sample. Evaluating the performance of processing approaches on flood forecasts is critical, since we can expect the

21    processing approaches to be less efficient for extreme and often unique flood events. Using a large catchment sample

22    allows us to investigate how the performance depends on both climatological and physiographic catchment

23    characteristic and to draw more robust conclusions. Furthermore, this comprehensive evaluation is performed for lead

24    times ranging from 1 to 9 days and the performance is assessed for different seasons.

25    Following the works cited above, the working hypothesis of this paper is that pre- and/or postprocessing improves

26    streamflow forecasts, but that the improvement might differ between catchments and between events. The main

27    objective of this study is to assess the potential improvements in flood forecasts by combining pre- and postprocessing

28    for a variety of catchments. We addressed the following questions:

29      1.   Which pre- and postprocessing approaches should be used in the hydrometeorological forecasting chain to

30        improve streamflow forecasts with an emphasis on flood forecasting?

31      2.   Are there regional or seasonal patterns in preferred pre- and postprocessing approaches?

32    In this study, we applied and evaluated the different processing schemes within the operational flood forecasting setup

33    used by the Norwegian flood forecasting service. The different schemes were tested for 119 catchments that vary in

34    climatology, catchment characteristics, and hydrological regimes. The large number of flood events and catchments

35    allowed us to provide robust assessments of the performance of the different schemes under different flood conditions.

1 ## 2 Study Area, Hydrological Model and Data

2 ### 2.1 Area

3 The west coast of Norway forms a topographical barrier for the westerlies. The resulting orographic enhancement of
4 precipitation makes this area one of the wettest parts of Europe, with an annual precipitation of around 4000mm,
5 whereas the driest regions in the rain shadow of the mountains have annual precipitation of around 400mm (Hanssen-
6 Bauer, 2017). The temperature depends both on latitude, altitude, and distance from the coast. The catchments belong
7 to Köppen-Geiger climate classes ranging from subarctic in the north and at high elevations, to temperate in the coastal
8 areas (according to the Köppen-Geiger climate classes as defined in Peel et al., 2007).

9 The spatial patterns of mean precipitation explain most of the spatial patterns in mean runoff. The seasonal variation
10 in runoff depends on seasonal variations in both temperature and precipitation. There are two basic runoff regimes in
11 Norway. For coastal regions with a temperate climate, the highest flows occur during autumn and winter due to heavy
12 rainfall. For inland regions with a sub-arctic or arctic climate, prolonged periods of winter temperatures below zero
13 ºC result in a seasonal snow storage, winter low flow, and high streamflow during spring due to snowmelt. There are,
14 however, many possible transitions between these two basic patterns (e.g., Gottschalk et al., 1979).

15 The study area consists of 119 catchments distributed all over Norway (Fig 1). All selected catchments are part of the
16 operational flood forecasting system and are mostly unregulated, with a large variation in size (3 to 15447 km$^2$) and
17 elevation (103 to 2284 meter above sea level [m.a.s.l.]). Six catchments are presented in more detail, the location of
18 these are indicated in Fig 1 and some key characteristics in table 1. The three first catchments are used as examples of
19 changes in reliability, depending on processing methods, datasets, and lead time. The three last catchments are used
20 to illustrate streamflow forecasts estimated by different processing approaches for three different flood events.

21 ### 2.2 Hydrological Model

22 We used the Hydrologiska Byråens Vattenbalance (HBV) model (Bergstrøm, 1974; Beldring, 2006; Sælthun, 1996)
23 that is used in the operational flood forecasting service at the Norwegian Water resources and Energy Directorate
24 (NVE). The HBV model is a conceptual model where the vertical structure of the model includes a snow routine, a
25 soil moisture routine, and a response function that consists of two tanks. Quick runoff is represented by a non-linear
26 tank, whereas slow runoff is represented by a linear tank. The model divides each catchment into 10 elevation zones
27 where each represents 10% of the catchment area. Catchment average temperature and precipitation are elevation
28 adjusted using a catchment specific lapse-rate to attain one representative precipitation and temperature value for each
29 elevation zone. The Nash-Sutcliff efficiency (Nash and Sutcliffe, 1970) and volume bias are used as calibration metrics.
30 The calibration period, 1996-2012, gives a mean Nash-Sutcliffe 0.77 for all 119 catchments, with zero volume bias.
31 The validation period, 1980-1995, shows mean Nash-Sutcliffe 0.73, with a mean volume bias of 5% (Ruan, 2016).

32 ### 2.3 Data

33 #### 2.3.1 Meteorological observation SeNorge v1.1

34 We used the gridded daily temperature and precipitation data from SeNorge v 1.1 that covers all of Norway with a 1x1
35 km grid size. The interpolation of observations to the grid is based on measured values at approximately 400

1  meteorological stations for precipitation, and 240 stations for temperature. Residual kriging is used for spatial

2  interpolation of de-trended temperature values (Tveito, 2007; Mohr, 2008). Temperature is detrended by adjusting

3  station data to sea level using a standard temperature lapse rate of 0.65 °C/100m. Triangulation is used for the spatial

4  interpolation of precipitation (Tveito, 2007; Mohr, 2008). The precipitation is further elevation corrected, using a

5  constant increase of 10% per 100 m beneath 1000 m.a.s.l, and 5% per 100 m above 1000 m.a.s.l. (Tveito et al., 2005).

6  **2.3.2   Meteorological forecasts ECMWF ENS.**

7  The temperature and precipitation forecasts used in the hydrological simulations of this study were taken from the

8  European Center of Medium-Range Weather Forecast (ECMWF) forecast ensembles (ENS). ENS provides an

9  ensemble of 51 members, with a forecasting period of 246 hours. The generation of the members of the ensemble is

10  done by adding small perturbations, which represent the uncertainty in the observations, to the forecast initial

11  conditions. Further, the uncertainty associated with the model physics is represented by perturbing the physics

12  tendencies that come from the parametrizations and each member is perturbed individually. This method is known as

13  the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme and improves the forecasts giving a much

14  better spread-error relationship compared to initial condition perturbations alone. A detailed description of the

15  ECMWF ENS system is provided in e.g. Buizza et al. (1999) and Persson (2015). The grid resolution of the model

16  forecasts used implemented in this study is 0.25° (i.e. model cycles/versions 40r1, and 41r1 (ECMWF, 2018b)). The

17  variables used for the hydrological modelling are the 2-meter temperature and the accumulated precipitation

18  aggregated to catchment daily (06:00-06:00) mean values.

19  **2.3.3   Streamflow reference simulations**

20  The streamflow measurements from the NVE database (https://www.nve.no/hydrology/ ) were used as a reference for

21  the hydrological model calibration. To evaluate the streamflow forecasts, we used simulated streamflow created by

22  running the hydrological model with SeNorge temperature and precipitation as forcing. Using this approach, we

23  isolated the effect of the uncertainty in the weather forecasts, and we could ignore uncertainty in hydrological model

24  parameters, parametrizations, and calibration.

25  **2.4   Study period**

26  The years 2014 and 2015 were chosen as the study period since several large floods affected rivers in most parts of the

27  country during this two-year period (Figure 1). In May 2014 there were large snowmelt floods in central and eastern

28  parts of Norway (affecting the Lågen, Glomma, and especially the unregulated Trysilelva catchments). In October

29  2014 western Norway was hit by an atmospheric river (a narrow plume of high moisture content transported from the

30  tropical and extratropical latitude towards the poles, see e.g., Zhu and Newell 1998), which led to flooding of multiple

31  rivers. Atmospheric rivers are responsible for extreme precipitation events when the moist air masses are

32  orographically lifted at topographical barriers like the west coast of Norway (e.g., Stohl et al., 2008). In July 2015 there

33  were snowmelt floods in Oppland (central eastern Norway), and in September 2015 an extratropical cyclone, *Petra*,

34  caused floods in Southern Norway. In early October 2015, a cyclone, *Roar*, that caused floods in Trøndelag and

35  Nordland and in early December a cyclone, *Synne*, caused floods in several catchments in south-west Norway, some

36  exceeding the 200-year return level.

1  During the study period 2014 and 2015, floods did not occur in all catchments; hence, the number of catchments used

2  in the flood evaluation analysis was reduced to 80. We still used all 119 catchments when evaluating the performance

3  for all streamflow values.

4  **3    Pre- and postprocessing**

5  We applied processing steps to both the weather input to the hydrological model and its streamflow output. To

6  distinguish the different processing steps, we refer to preprocessing as corrections schemes applied to temperature and

7  precipitation ensembles, and postprocessing as corrections applied to the hydrological ensembles.

8  **3.1    Processing chain**

9  The temperature and precipitation forecast data from ECMWF were prepared by aggregating the variables from hourly

10  to a daily time step. Thereafter the horizontal resolution was changed using nearest neighbor interpolation to a 1×1 km

11  grid, equal to the SeNorge grid. For the temperature forecasts, a standard elevation adjustment of 0.65°C/100m was

12  applied to account for the elevation differences between the original and the seNorge grid. Finally, the temperature

13  and precipitation forecasts were aggregated to average values for each catchment. We used the ECMWF forecasts from

14  2014 and 2015 to force the hydrological model, which enabled a retrospective evaluation of the daily streamflow

15  forecasts for almost two years. The unprocessed daily forecasts for each catchment are referred to as $Traw_{t,l,s,m}$ and

16  $Praw_{t,l,s,m}$ where $t$ is issue time, $l$ is lead time, $s$ is catchment and $m$ is ensemble member. For temperature and

17  precipitation forecasts, two different preprocessing approaches were chosen, a grid calibration (CAL) producing the

18  ensembles $Tcal_{t,l,s,m}$ and $Pcal_{t,l,s,m}$, and Bayesian model averaging (BMA) producing the ensembles $Tbma_{t,l,s,m}$ and

19  $Pbma_{t,l,s,m}$. For postprocessing of streamflow, we used BMA to create $Qbma_{t,l,s,m}$. For all approaches, the processing

20  was applied to each issue date, $t$, lead time $l$ and catchment, $s$, independently. To improve readability, $t,l,s,m$ is

21  suppressed in the remainder of this paper. We evaluated all combinations of $Tcal$ and $Pcal$ together with $Traw$ and

22  $Praw$, as well as all combinations of $Tbma$ and $Pbma$ together with $Traw$ and $Praw$. $Tcal$ and $Pcal$ was not combined

23  with $Tbma$ and $Pbma$. The seven combinations of temperature and precipitation were run through the hydrological

24  model resulting in seven unprocessed streamflow forecasts ($Qraw$). Thereafter, postprocessing the raw forecasts

25  resulted in seven streamflow forecasts ($Qbma$), which could be compared to $Qraw$ to establish the effect of

26  postprocessing. Figure 2 provides an overview of the complete processing chain. More detailed presentation of each

27  step in the processing chain follows.

28  Different observational reference data and periods were the basis for the different processing techniques. An overview

29  of the variables, resolution, and data used for training are presented in Table 2 and details are provided in the following

30  subsections.

31  **3.2    Grid calibration**

32  The Norwegian Meteorological Institute (MET Norway) uses grid calibration approaches to improve ensemble

33  forecasts that are used for the operational national weather forecasts published at yr.no (methods available at

34  https://github.com/metno/gridpp/). We have rerun the preprocessing of the daily ensemble forecasts of temperature

1 and precipitation between 2014 and 2015, using the operational processing methods at that time. In the following text
2 these are referred to with the subscript *cal*. All calibration parameters were provided by MET Norway.

3 For the grid calibration methods, we applied the same corrections to each ensemble member. The ordering of members
4 was therefore kept. Thereby consistency between the calibrated temperature and precipitation members was ensured
5 and the temporal profile was preserved, which is important for the hydrological modelling.

### 3.2.1 Temperature calibration (Tcal)

7 Quantile mapping (Seierstad, 2016; Bremnes, 2007) was used to remove biases in the temperature forecasts by moving
8 the ensemble (ENS) forecast climatology closer to the observed climatology. MET Norway used Hirlam (Bengtsson
9 et al., 2017) temperature forecast at a $4\times4$ km$^2$ grid, as a reference for parameter estimation used to calibrate the
10 ECMWF ENS. Hirlam was the operational regional model at the time and is suitable as a reference since it provides a
11 continuous field covering all of Norway at a sub daily time step. Hirlam gives a higher skill and is less biased than
12 ENS when they are compared to point observations (Engdahl et al., 2015). To establish the calibration parameters,
13 MET Norway used both ENS reforecasts (Owens, 2018) and Hirlam data from July 2006 to December 2011
14 interpolated to a $5\times5$ km$^2$ grid. The ENS reforecast is a 5-member ensemble generated from the same model cycle
15 (40r1 and 41r1) as the operational ENS forecasts. For each grid cell, quantile transformation coefficients unique for
16 each month of the year, were determined by using data from a three-month window centered on the target month, e.g.
17 the May analysis consists of April, May, and June (Seierstad, 2017). The coefficients were estimated by mapping the
18 first 24 hours of the forecasts. A 1:1 extrapolation was used for forecasts outside the range of observation. In this study
19 we used the quantile transformation coefficients estimated by MET Norway. This enabled us to establish a
20 retrospective calibration of the temperature ensemble forecasts.

### 3.2.2 Precipitation calibration (Pcal)

22 To account for the intermittent nature of daily precipitation, a Bernoulli-Gamma distribution was used to calibrate the
23 precipitation forecasts. Precipitation observations from around 200 WMO stations in Norway are used to establish the
24 parameters of the Bernoulli-Gamma model. All parameters in the Bernoulli-Gamma model depend on lead-time, but
25 independent of location and issue date.

26 The probability mass for zero precipitation was specified by logistic regression. Both the cube transformed, and the
27 untransformed ensemble means and the fraction of ensemble members with precipitation higher than 0.5mm were used
28 as predictors. A total of four parameters were estimated in the logistic regression model. The precipitation amounts
29 were modelled by a gamma distribution The cube root of the forecast ensemble mean is used as a predictor in a model
30 with two parameters to fit the mean, whereas the untransformed forecast ensemble mean is used as a predictor in a
31 model with two parameters are used to fit the standard deviation. MET Norway provided the parameters that were used
32 at the issue time of the precipitation forecasts, and we applied them for a retrospective calibration of the precipitation
33 ensemble forecasts.

### 3.3 Bayesian Model averaging

1    Bayesian model averaging (BMA) aims to correct dispersion errors in a bias corrected ensemble (Raferty et al 2005).

2    For each lead time, BMA uses a mixture distribution, where for an ensemble with $M$ members, the density function

3    conditioned on all ensemble members is the weighted average of kernels for each member $m$. The preprocessed

4    meteorological ensembles were established by randomly drawing $M$ realizations from the mixture distribution

5    estimated by BMA. The kernel, for the quantity one wishes to forecast, $y$, is denoted by $f_\theta(y|x_m)$ where $x_m$ is the raw

6    forecast's ensemble member $m$ and $\theta$ are parameters of the kernel pdf $f$. The probability density function conditioned

7    on all $M$ ensemble members is the weighted average of the pdf for each member:

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

8    where $\sum_{m=1}^{M} w_m = 1$ and the weights are interpreted as the posterior probabilities of each ensemble member. The

9    ensembles in this paper are based on ECMWF ENS which are considered exchangeable, and weights and parameters

10   can be constrained to be equal for all members (Fraley et al 2010). For each issue date we used the previous $n$ days of

11   ensemble forecasts and reference observations to estimate the parameters in the kernel. To account for the specific

12   properties of temperature, precipitation and streamflow, different kernel distributions were used, the details are

13   provided below.

### 3.3.1    BMA for temperature (Tbma)

15   We followed Raferty et al (2005) and used a Normal distribution as the kernel for the temperature BMA models. Since

16   the temperature ensemble forecasts were not already bias corrected, the mean is specified as $a_0 + a_1 T_{raw,m}$ , where

17   $T_{raw,m}$ is the temperature forecast for ensemble member $m$ and $a_0$ and $a_1$ are regression parameters that account for any

18   bias. The parameters are specific for each catchment, issue date and lead time and are the same for all ensemble

19   members.

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

20   To estimate the parameters, the catchment average temperatures from SeNorge were used as a reference.

### 3.3.2    BMA for precipitation (Pbma)

22   We followed Sloughter et al (2007) who proposed a Bernoulli-gamma distribution as kernel in the BMA precipitation

23   models to establish $P_{bma}$.

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

24   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

25   of zero precipitation given by a logistic regression model:

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

1    where $b_0$, $b_1$ and $b_2$ are regression parameters common for all ensemble members and $\delta_m$ equals 1 if $x_m = 0$ and equal

2    0 otherwise.

3    $h(P_{bma}|P_{raw,m})$ was assumed to follow a gamma distribution for the cube root transformation $P'_{bma} = P_{bma}^{1/3}$ of the

4    precipitation, where the mean ($\mu_m$) and variance ($\sigma_m^2$) of the distribution depend on the ensemble member:

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

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

6    Bernoulli-gamma kernels were estimated using the catchment average precipitation from seNorge as reference.

7    **3.3.3    BMA for streamflow (Qbma)**

8    We applied a Box-Cox transformation (Box and Cox 1964; e.g., Duan et al. 2007) on both observed and forecasted

9    streamflow to make the transformed streamflow $q^*$ normally distributed:

$$q^* = \begin{cases} \dfrac{(q^\lambda - 1)}{\lambda} & for \ \lambda \neq 0 \\ log(q) & for \ \lambda = 0 \end{cases} \tag{6}$$

10   here $\lambda$ is a transformation parameter. The Box-Cox transformation has proven valuable for hydrological applications

11   (e.g., Engeland et al. 2010; Bates and Campbell 2001; Thyer et al. 2002; Yang et al 2007). We used a fixed $\lambda$ based on

12   previous studies by Engeland et al (2010), who found that $\lambda = 0.2$ gave forecast errors that were approximately

13   independent of forecasted values. As for temperature, we applied the BMA with a mixture of normal kernels for

14   postprocessing the streamflow forecasts.

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

15

16   **3.4    BMA training length**

17   Following Raferty et al. (2005), the BMA models for temperature, precipitation and streamflow were trained on data

18   from a time window prior to the issue date for each forecast. We tested different training lengths for all variables and

19   lead times, using CRPS (description in following section) as evaluation metric. Experiments with different training

20   lengths showed that the optimal window size depends on variable, lead-time, and whether CRPS was calculated for all

21   data or only for days with flooding (example in Fig 3). Precipitation was most sensitive to the training length due to

22   the necessity of precipitation occurring within the time window. 45 days training period was optimal for most

23   catchments and lead-times (A-Fig 1 and 2). To keep a consistency during the evaluation we used 45 days training

24   period for all variables (i.e., temperature, precipitation, and streamflow).

25   **3.5    Temperature and precipitation dependence structure (Ensemble copula coupling)**

26   The BMA models described above were applied independently to each weather variable, each location (here

27   catchment) and each lead time. The preprocessed ensembles where established by drawing 51 new realizations from

1    the mixture distribution of each BMA model independently. To recreate forecast trajectories of temperature and

2    precipitation, it is necessary to account for the temporal and inter-variable dependence structures. In this study, it was

3    achieved by using an approach similar to Ensemble Copula Coupling (ECC, Schefizik et al., 2013). The original 51

4    ensemble members ($m$) for temperature and precipitation were, for each location, issue date, and lead time, assigned a

5    rank ($r_{o,m}$). Similarly, the 51 BMA-processed precipitation and temperature ensemble members were assigned a rank

6    ($r_{n,m}$). The 51 preprocessed ensemble members were reordered by using $r_{o,m}$ and $r_{n,m}$ as keys to keep the preprocessed

7    ensemble in the same rank sequence as the original ensemble members. By applying this method to all variables, lead

8    times, and issue dates we maintain the dependency between the variables, as well as the temporal dependency for each

9    of the variables.

10    **4    Evaluation**

11    We evaluated the pre- and postprocessing methods for all days of the study period using the complete dataset, as well

12    as for the flood dataset.

13    **4.1    Reliability: Cumulative rank-histogram plots**

14    The reliability of an ensemble forecast is often visually presented by the rank-histograms (Anderson, 1996; Talagrand

15    et al., 1997; Hamill, 2001). In our setup, the rank-histograms consist of $i$=52 bins (51 members +1), where the value

16    of the ordered ensemble members defines the limit between the bins. Each bin in the rank-histogram reflects the

17    frequency of the ranked reference observations compared to the ensemble forecast, and a reliable forecast should have

18    a uniform distribution of observations between the bins. There are 14 rank-histograms for each lead time and catchment

19    to be evaluated. To reduce the number of plots, we evaluated the reliability by creating a Q-Q plot based on the

20    cumulative rank-histogram (scaled to unity) on the y-axis and the uniform distribution on the x-axis, as explained in

21    Fig 4. The cumulative rank-histogram $F_i$ for bin $i$ is the sum of the relative frequency $f_k$ for all bins where $F_i =$

22    $\sum_{k=1}^{i} f_k$. The expected relative frequency of observations in each of the 52 bins given a uniform distribution equals

23    1/52, represented by the cumulative uniform distribution $U_i = \sum_{k=1}^{i} \frac{k}{52}$. . In this cumulative rank-histogram plot the

24    1:1 line represents a uniform rank-histogram with an equal probability for the observations to be located within each

25    bin. This approach enabled us to compare the reliability for all 14 processing schemes within a single plot. The shape

26    of the cumulative rank histogram plots enables the detection of biases as well as under- and over dispersion as explained

27    in the Fig 4.

28    **4.2    Continuous rank probability score (CRPS) and - skill score (CRPSS)**

29    The continuous rank probability score (CRPS) has properties that are appealing for the evaluation of ensemble forecast.

30    Firstly, it is sensitive to the entire permissible range of parameters of interest. Secondly, its definition does not require

31    predefined classes, which might influence the results. For a deterministic forecast, CRPS reduces to the mean absolute

32    error (MAE, Hersbach, 2000), which enables a comparison between a deterministic and an ensemble forecast. CRPS

33    measures the integral of squared difference between the forecast and the observation, both given as cumulative

34    distribution function (cdf). If the observation is deterministic the Heaviside function is used for the observation cdf

35    (Hersbach, 2000). For ensemble forecasts, the CRPS is calculated discretely since both the observations and the

36    forecasts are reported in discrete intervals (Hersbach, 2000, Eq. 8):

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

1 Where $M$ is the ensemble size, $x_m$ is ensemble member $n$ and $x_{obs}$ is the reference observation. For a time-series of

2 forecasts, the mean CRPS for each scheme ($\overline{CRPS_{PS}}$) can be calculated. CRPS will give credit to high probabilities

3 close to the reference, which is not necessarily the case for other ensemble verification scores (Gneiting and Rafterty,

4 2007). CRPS has the same unit as the observations (m$^3$/s for streamflow), and is negatively oriented, where zero is the

5 optimal value.

6 The continuous ranked probability skill score (*CRPSS,* Eq. 9) enables assessment of the skill of the different processing

7 schemes (*PS*) relatively to the raw forecasts (raw). The mean CRPS for each scheme ($\overline{CRPS_{PS}}$) and for the unprocessed

8 forecasts ($\overline{CRPS_{raw}}$) are used to calculate CRPSS.

$$CRPSS_{PS} = 1 - \frac{\overline{CRPS_{PS}}}{\overline{CRPS_{raw}}} \tag{9}$$

9 Note that CRPSS has 1 as the optimal value and is positively oriented. Since CRPSS has no units, we could calculate

10 average skill scores across all catchments. *CRPS* and *CRPSS* were calculated for the complete dataset as well as well

11 as for the flood dataset.

12 **4.3    The Critical success index (CSI)**

13 In an operational flood forecasting setting, flood warnings are issued when there is a certain probability for streamflow

14 to exceed predefined flood warnings thresholds. The occurrence and non-occurrence of floods are therefore binary

15 events that can be summarized in a contingency table providing an overview of hits (H), missed events (M), false

16 alarms (F), and correct non-events (N). Based on the contingency table shown in Table 3, the following indices can be

17 used to evaluate the performance of a forecasting system.

18 Hit ratio, where a hit rate of 1 is the best performance (S$_R$):      $S_R = \frac{H}{H+M}$

19 False alarm ratio (F$_R$):      $F_R = \frac{F}{H+F}$

20 Critical Success Index (CSI):      $CSI = \frac{H}{H+F+M}$

21 Since floods are rare events, there is a small number of flood-events compared to the number of non-events.  A good

22 forecast has a high hit ratio and a low false alarm ratio. The Critical Success Index (CSI, Donaldson et al., 1975; Jolliffe

23 and Stephenson, 2018) balance these two aims by penalizing the hit ratio for both the missed events (M) and the false

24 alarms (F). In an operational setting, a warning will be issued when a predefined number of ensemble members (or a

25 defined probability) exceeds the flood warning threshold. The probability of exceedance opens for potential cost lost

26 evaluation, however for the simplicity of this work we have chosen a limit of 10 members exceeding the mean annual

27 flood level. The mean annual flood has of a return period of 2.33 years (i.e. ~20% probability of occurrence).

28 **4.4    Floods by seasons**

1  There might be several reasons for the seasonal differences in flood forecast performance. Firstly, there are biases in

2  forecasted temperatures, especially for the Norwegian coast during autumn and winter (Seierstad et al., 2016, Hegdahl

3  et al., 2019). Secondly, the flood-dominating processes are often aligned to different season, e.g. snowmelt contribution

4  to floods dominates in spring, and rain-induced floods dominate in autumn. For these reasons, we divided the flood

5  events into spring and autumn floods and used *CSI* to evaluate how the performance of processing methods depend on

6  season. The available data covers a period of two years and we defined spring from April 4 to June 13, and autumn

7  from September 01 to December 10. Both seasons consist of 2×101 days and 35 catchments were affected by spring

8  floods and 40 catchments by autumn floods.

9  **5    Results**

10  We assessed the reliability of the raw and processed streamflow forecasts, and results for selected catchments are

11  presented. *CRPS* and *CRPSS* were used to evaluate the different processing schemes for the full dataset and the flood

12  dataset. Furthermore, we evaluated the effect of pre- and postprocessing regarding location, by plotting maps of the

13  processing schemes giving the highest performance on the flood dataset. *CSI* was used to assess the ability to predict

14  the exceedance of flood warning levels for the different schemes. *CSI* was calculated for all floods as well as for spring

15  and autumn floods separately. Finally, we present streamflow forecasts based on  the different processing approaches

16  for three flood events.

17  **5.1    Reliability**

18  We used cumulative rank-histogram plots to compare all 14 processing schemes for all lead times and found that for

19  most catchments the schemes improved the reliability of the forecasts. Examples for lead times 1, 5 and 9 for three

20  catchments chosen to highlight some differences, are shown in Fig 5. Vaekkava (Fig 1, Table 1) is representative of

21  the effect of pre- and postprocessing for most catchments in this study. The raw ensembles (*Traw_Praw*) have a

22  negative bias for all lead times. For a lead time of 1 day, all postprocessing schemes produce reliable forecasts, whereas

23  preprocessed forecasts still underestimate the streamflow forecasts. The preprocessed forecasts become more reliable

24  with increasing lead time. This can be explained by an increasing spread in the ensemble for longer lead times. For a

25  lead time of 9 days, we see that the preprocessed forecasts, independent of methods, are more reliable than the

26  postprocessed forecasts. Refsvatn (Fig 5 second row, Table 1) has a slightly positive bias in the raw ensemble

27  (*Traw_Praw*) for a lead time of 1 day. The preprocessing schemes results in forecasts with a large negative bias and

28  hence makes the forecasts less reliable, whereas schemes with postprocessing (*\*_Qbma*) improve the reliability. For

29  lead times of 5 and 9 days, the raw ensembles are the most reliable. Tannsvatn (Fig 5 bottom row, Fig 1, Table 1) has

30  raw forecasts that are rather reliable for all lead times. For a lead time of 1 day, the improvements are seen by all

31  postprocessed ensembles whereas the preprocessing introduces a negative bias. For a lead time of 5 days, the reliability

32  is similar for most processing schemes, but poorest for the preprocessing schemes *Pcal* and the *Tbma*. At a lead time

33  of 9 days, however, the preprocessing schemes based on *Pbma*, performs best, while those that include postprocessing

34  are least reliable.

35  **5.2    Skill – relations to lead time for all data and floods**

1 We used CRPS and CRPSS to evaluate how the different processing methods affected the performance of ensemble

2 streamflow forecasts for all lead times and catchments. In Fig 6 the *CRPSS* for all data and catchments is presented.

3 The most striking finding is that nearly all catchments benefit from processing. Postprocessing in combination with

4 preprocessing is most important for the short lead times. *Pcal* show the largest variability in performance, where a

5 larger portion of catchments only slightly benefit from *Pcal*, indicating that this preprocessing is the least robust. For

6 the flood dataset (Fig 7), there is a larger difference between the median of *CRPSS* for the schemes compared to Fig

7 6. However, the variability in skill is larger for the flood dataset compared to the full dataset, meaning that there are

8 fewer catchments benefiting from the processing schemes under flood conditions. Postprocessing without

9 preprocessing seems to be the least good approach. For the longer lead times, there are increasingly more catchments

10 where postprocessing leads to a poorer performance, compared to using the raw forecast.

11 Additional results are shown in A-Fig 3 and 4 for the full dataset and A-Fig 5 and 6 for the flood dataset, all these

12 figures are in the appendix. By only focusing on the best processing approach for the single catchments, the applied

13 postprocessing methods are most important for the short lead times (1-3 days) when analyzing the complete dataset

14 (seen by the yellow to green colors in A-Fig 3 and supported by the histograms in A-Fig 4). We moreover find that the

15 most skillful method can change for a catchment with lead-time (A-Fig 3). The BMA applied to temperature and in

16 the combination of BMA applied to precipitation are the two best methods for lead-times above 3 days.

17 For the flood dataset we find that there are no systematic patterns to whether pre- or postprocessing is most important

18 to improve the skill (A-Fig 5). Postprocessing performs similar to preprocessing for most lead-times and is hence less

19 important for the short lead-times compared to what was found for the full dataset. BMA seems to be the better choice

20 for preprocessing, and improves the performance for more catchments compared to CAL. For longer lead times, BMA

21 on temperature is the most important method for improved *CRPS*. The general tendency seems to be that preprocessing

22 precipitation is most important for the short lead-times, whereas preprocessing temperature is more important for the

23 longer lead-times.

24 Figure 8 gives a detailed presentation on how the mean *CRPS* varies with lead time, processing scheme, and the

25 evaluation dataset for three individual catchments. For the full dataset (Fig 8 left), the *CRPS* for postprocessed forecasts

26 increases faster with lead time than *CRPS* for forecasts without postprocessing. The lead time at which postprocessing

27 gives better performance than not using postprocessing varies between catchments. This is supported by the results

28 presented in A-Fig 3. A striking difference is that *CRPS* increases with lead time when the full dataset is used, whereas

29 it is reduced by lead time for the flood dataset (Fig 8 right) for several of the processing schemes. The pattern for the

30 full dataset (i.e. *CRPS* increases with lead time) is representative for most catchments, whereas changes in *CRPS* with

31 lead time for the flood dataset varies between the catchments. We see that the mean *CRPS* for all streamflows (Fig 8

32 left) is smaller than for floods (Fig 8 right), which can be explained by the data used to estimate the mean. The flood

33 dataset consists of fewer days and higher values, and hence the possibility for larger errors. An explanation for the

34 decrease in *CRPS* for the flood dataset in Fig 8 right is that the ensemble spread increases with lead time, and it is

35 therefore more likely that the observed floods are within the ensemble range for the long lead times.

36 **5.3    Skill – relations to location**

1    Figure 9 shows a map of which processing method that achieves the highest performance according to *CRPS* for the
2    flood dataset for each catchment for lead time 1, 5, and 9 days. The left column shows whether a preprocessing scheme
3    alone or a combination of pre- and postprocessing methods gives the highest performance. The figures show that
4    inland, high elevation, and eastern catchments are improved by postprocessing for lead times of 1 and 5 days, whereas
5    the coastal catchments do not attain the highest score by postprocessing. In the right column we show which of the
6    BMA preprocessing approaches that resulted in the best *CRPS*. Catchments where the grid-calibration or the raw
7    forecasts gave the best performance are shown as black dots. We find that *Pbma*, alone or in combination with *Tbma*,
8    gives the best results for western and southern coast of Norway for lead times of 1 and 5 days. For a lead time of 9
9    days, however, *Tbma* alone is more important. In the coastal regions, floods are mainly rain driven, and we find that
10   *Pbma* performs well in these regions. BMA on temperature alone has a less clear pattern. A summary of the numbers
11   from Fig 9 is presented in table 4 and quantifies the visual information from Fig 9. The effect of postprocessing is
12   larger for shorter lead time and the catchments where preprocessing was the best option, BMA is the best choice for
13   about 70 to 80 % of the catchments. Combining *Tbma* and *Pbma* performs best for a larger group of catchments.

14   **5.4    CSI for the whole year, spring, and autumn floods**

15   In this evaluation, the processing scheme giving the highest CSI for each catchment is considered, and we counted the
16   number of catchments for which the specific scheme gave the best CSI. For each catchment, multiple methods can
17   achieve equal CSI. Therefore, for some lead times, the number of "best" CSI exceeds the total number of catchments.

18   We first evaluated CSI for floods from the whole year (A-Fig 8), which did not give any clear indications of methods
19   that performed better than others. However, by separating the flood dataset between floods occurring in spring (Fig.
20   10) and those occurring in autumn (Fig 11) we attain some interesting insight. For spring (Fig 10) most methods give
21   good results for multiple catchments, indicating more than one successful method. The improved predictions by
22   applying pre- and/or postprocessing to spring floods, holds for most lead times. For lead times of 2 to 5 days
23   postprocessing provides the best CSI for more catchments than preprocessing alone, whereas beyond 5 days' lead time
24   we find that about half of the successful predictions includes postprocessing.

25   For autumn (Fig 11) the results diverge from the spring results. For a lead time of 1 day, the predictions are highly
26   improved by including postprocessing, whereas the effect of postprocessing diminish for lead times of 2 and 3 days.
27   From a lead time of 4 days there is no predictability by most methods, and only six catchments show predictive skill
28   by applying *Tbma* alone or in combination with *Pbma*.

29   **5.5    The effect of pre- and postprocessing for a selection of events and catchments**

30   The forecasted streamflow is essential to determine a correct flood warning level. In this subsection we present three
31   flood events and catchments to exemplify how the different processing approaches influences the ensemble flood
32   forecasts. The events are the atmospheric river affecting western Norway in October 2014, the extreme weather event
33   *Synne* hitting southern Norway in early December 2015, and a snowmelt flood in eastern Norway in May 2014. For
34   all examples, the issue date of the forecast is selected 3 to 5 days before the peak of the flood.

35   Figure 12 shows the outcome of the different processing approaches for the October 2014 event at Bulken (Fig 1, Table
36   1) in western Norway. Some of the ensemble members reach the reference streamflow (black line) when *Pbma* is

1 applied without *Qbma*. However, none of the ensemble medians reach up to the level of the reference streamflow

2 (black line). *Pbma* induces very high streamflow for some of the members, whereas BMA applied to streamflow

3 removes the effect of *Pbma* (Fig 12 left and right respectively). The large spread in streamflow when using *Pbma*

4 indicates large uncertainty in the precipitation forecasts for this event.

5 The extreme weather event in December 2015 was difficult to forecast. In particular, the location of the rainfall was

6 highly uncertain. Figure 13 shows the outcome of the different processing approaches for this event at Moeska (Fig 1,

7 Table 1) in south-western Norway. We see that precipitation is underestimated, and none of the processing schemes

8 result in ensemble members that reach the reference level for streamflow. For this event at Moeska the same pattern is

9 seen as for the event at Bulken, where *Pbma* induces high streamflow values (Fig 13 left) that are later suppressed by

10 the *Qbma* (Fig 13 right).

11 Figure 14 shows the outcome of the different processing approaches for the snowmelt flood in May 2014 at

12 Nybergsund in eastern Norway. This flood is best forecasted by the raw and preprocessed input, with small differences

13 between the schemes. Postprocessing reduces the median forecasts for all lead times, in addition to increasing the

14 spread.

15 **6  Discussion**

16 The results demonstrate that all catchments benefitted from one or more of the applied processing schemes, thereby

17 confirming our working hypothesis. However, it was not possible to identify a distinct processing chain that was

18 optimal for all forecasts, the choice of method depends on several factors including lead time, season, location, and

19 evaluation criteria.

20 A part of the answer to our first research question "Which pre- and postprocessing approaches should be used in the

21 hydrometeorological forecasting chain to improve streamflow forecasts with emphasis for flood forecasting?" is that

22 preprocessing using catchment specific BMA generally performed better than the gridded calibration (CAL). One

23 explanation is that the BMA calibration uses the same temperature and precipitation data that were used to tune the

24 hydrological models and establish the reference streamflow. Using grid-calibrated temperature and precipitation might

25 therefore, in many cases, lead to biases in streamflow forecasts. One example is Refsvatn (Fig 5 LT: 1) where the CAL

26 methods induce a larger bias compared to the BMA methods. Another aspect is that the BMA approaches tailor the

27 preprocessing to each catchment, whereas the model for the grid calibrated precipitation is independent of location and

28 is therefore less flexible (See Table 2). In Fig 6 we see a large variability in performance for *Pcal*. Even though *Pcal*

29 performs well for a majority of the catchment when considering the full dataset, several catchments show only small

30 or no improvement to the forecast skill. Postprocessing, i.e. combining *Pcal* and *Qbma*, assists in improving the

31 forecasts for these catchments.

32 It is moreover instructive to see that postprocessing alone seems to be the least optimal choice when evaluating both

33 the full dataset and even less optimal when the subset of floods is considered. This demonstrates the importance of

34 correcting biases and spread in the forcing. The catchments' responses to the temperature and precipitation inputs are

35 non-linear, in particular for snow accumulation and snow melt processes where temperature thresholds are important.

36 Using postprocessing alone is therefore less effective in correcting for biases in inputs to the hydrological model.

1   The combination of pre- and postprocessing approaches that outperforms the others depends on catchment, lead time,
2   streamflow magnitude, and the choice of evaluation metric. We find that for the complete dataset, the best CRPS is
3   seen when applying postprocessing combined with BMA preprocessing of temperature for lead times of up to three
4   days, whereas for the longer lead times BMA preprocessing of temperature alone or both precipitation and temperature
5   provide the best performance (Fig 6 and A-Fig 4). This result is in line with Benninga et al (2017) who underlines the
6   importance of improving the meteorological inputs, in particular for high flow events. Global meteorological
7   ensembles often lack spread for shorter lead times since they are designed for medium range forecasts and therefore
8   use perturbations that optimize the ensemble spread for longer lead times. BMA models used both for pre- and
9   postprocessing will therefor improve the forecast skill. It would be instructive to assess whether using regional
10  meteorological ensembles, which are better able to model the forecasts uncertainties in the short range compared to
11  their global counterparts (Frogner et al 2019a, 2019b), as inputs to the hydrological model alter this finding. However,
12  such forecasts were not available for our study period, but may be the focus of future research.

13  Comparing CRPSS in Fig 6 and 7, we see that the improvement in skill resulting from the processing schemes is
14  smaller for the flood dataset compared to the complete dataset. Looking at CRPS for the full dataset and floods (A-Fig
15  5 and 6 respectively) it is less evident whether any schemes outperform others for the floods whereas for full dataset
16  we see similar results as for CRPSS. We see that postprocessing is less useful for the three first lead times for the flood
17  dataset as compared to the full dataset. Using BMA for both precipitation and temperature for the shortest lead times
18  and only temperature for the longest lead times was the best choice for the largest portion of the catchments. In addition
19  to the differences in preferred processing schemes between catchments, we find that for a single catchment, the best
20  processing schemes varies depending on lead-time. This underlines that forecast errors arise from different sources,
21  and that being conclusive based on relatively small sample of floods is difficult.

22  In answer to our second research question "*Are there regional or seasonal patterns in preferred pre- and
23  postprocessing approaches?*" we found that the performance of the processing schemes has both regional and seasonal
24  patterns, when the flood dataset is used for evaluation. The regional pattern indicates that an excess of catchments
25  benefitting from preprocessing are located in coastal areas (Fig 9). Another finding is that those improved by BMA
26  applied to precipitation (*Pbma*) are in areas with high precipitation (the west and southwest coast of Norway, Fig 9).
27  It is also clear that *Tbma_Pbma* is the combination with the highest performance for a lead time of 1 day, with the
28  performance diminishing with lead time, and for a lead time of 9 days, *Tbma_Praw* is a better choice (Table 4 and Fig
29  9). Postprocessing is more important for the inland and high elevation catchments, where temperature and slower
30  snowmelt processes are dominating. Moreover, for these regions we see that the effect of postprocessing is smaller
31  with increasing lead time.

32  The seasonal effect was evaluated by separating spring floods from autumn floods. The CSI shows that there are large
33  differences in predictability between seasons. There is almost no ability to predict autumn floods beyond 3 days, only
34  for 6 of 40 catchments are floods predicted by any of the approaches. In contrast, the forecasts for the spring floods
35  show a predictability up to 9 days, and for 23 of the 35 catchments one or more approaches were able to predict the
36  floods. These results indicate that the predictability of floods depends on flood-generating processes, i.e. snowmelt
37  induced spring floods are easier to forecast than rain induced autumn floods. These results further imply that the autumn
38  precipitation and floods are the most difficult to predict and has the highest potential for improvements.

1    For some catchments we see contradictory results when comparing CRPS and CSI for the flood dataset. *Tbma* produces

2    the best CRPS for most catchments for longer lead-times (A-Fig 6), however *Tbma* gives a lower CSI compared to the

3    other preprocessing methods (A-Fig 7 and Fig 10-11). This indicates that care must be taken when choosing an

4    appropriate evaluation metric. CRPS indicates the error between the forecast and the reference value and favors

5    forecasts close to the reference (Gneiting and Rafterty, 2007). CSI on the other hand gives no favor for forecasts close

6    to the reference only to whether the forecast exceeds the warning threshold or not. For example, the processing scheme

7    that had the best CRPS might slightly underestimate the reference value, and if the reference is just above the warning

8    threshold, this scheme will miss the event, resulting in a low CSI value In contrast, a processing scheme that highly

9    overestimate the reference will result in a poor CRPS and a good CSI.

10    For the calculation of CSI, we used a limit of 10 ensemble members (a probability of about 20%) exceeding the flood

11    threshold to issue a flood warning. The ensemble can provide a whole range of probabilities and here we only evaluated

12    for one probability level. The optimal probability of exceedance to issue a flood warning might be different between

13    catchments, lead times, and seasons. Another aspect is to investigate the acceptance level for false alarms to missed

14    events. The number of tolerable false alarms might depend on the impacts of the event (e.g. risk evaluation), and it is

15    therefore difficult to make one absolute decision on behalf of all possible exceedance levels (flood sizes) and affected

16    parties. We acknowledge that the choice of evaluation criteria can be different depending on the users and the cost of

17    mitigation action compared to the loss due to an event, and that false alarms and missed events might be weighted

18    different depending on a total cost-loss evaluation.

19    One concern when using BMA for preprocessing precipitation is that some of the ensemble members in *Pbma* attained

20    physically non-plausible values. resulting in very high flood forecasts. This is apparent for the Bulken catchment for

21    the October 2014 event (Fig 12). This suggests that the forecast distribution can be sensitive to large errors in

22    precipitation. Especially for Western Norway where a steep topography causes large spatial differences in precipitation

23    and therefore a potential for large errors in forecasts, *Pbma* should be used with care. The region experienced large

24    amounts of precipitation prior to the October 2014 event. Therefore, the estimated BMA parameters are based on data

25    for a period with possible large errors in the forecasted precipitation, implicating large uncertainty in the BMA model

26    parameters. Possible solutions could be to use categorized approached (e.g., Ji et al., 2018), where the precipitation is

27    separated into precipitation categories (based on for example daily ensemble mean) and unique BMA models are

28    trained for each category.

29    **7    Conclusions**

30    In this study, we have evaluated streamflow forecasts in 119 catchments based on fourteen schemes with different

31    combinations of the raw, pre-, and postprocessed values. The modelling chain is similar to the operational flood

32    forecasting system, and we evaluated the forecast with a special emphasis on flood values exceeding the mean annual

33    flood (QM). From the results presented and discussed in this paper, we conclude that:

34    Applying pre- or postprocessing schemes improve streamflow forecasts compared to using raw forecasts. The best

35    combination of pre- and postprocessing approaches depends on location, season, lead time, and the purpose of the

36    forecasting as represented by different evaluation criterions. The large number of catchments used for evaluation

1 allows us to draw some general conclusions that can assist us in choosing an appropriate processing chain and to
2 identify which forecasts that are the most challenging.

3 *Which pre- and postprocessing approaches should be used in the hydrometeorological forecasting chain to improve*
4 *streamflow forecasts with emphasis for flood forecasting?*

5 • An evaluation of CRPS for the complete dataset of two years showed that the combination of pre- and
6 postprocessing is most effective for short lead times, up to two-three days. For longer lead times, processing
7 schemes that only include preprocessing provide the best results. BMA is the preferred method for
8 preprocessing, either applied to temperature (*Tbma*) alone or in combination with precipitation (*Pbma*).

9 • For days where floods exceeded QM the added value of processing is less clear. For a small majority of the
10 catchments applying BMA to precipitation and/or temperature (for longer lead times) improves the CRPS
11 compared to the raw forecast and is also better than grid calibration.

12 *Are there regional or seasonal patterns in preferred pre- and postprocessing approaches?*

13 • The processing is sensitive to regional or seasonal patterns. Postprocessing was most effective for inland and
14 higher elevated catchments. The coastal catchments gained more from preprocessing. Especially BMA
15 applied to precipitation and temperature improved CRPS for the western and southwestern coastal catchments
16 for the early lead times, whereas *Tbma* was most important for the longer lead times.

17 • The added value of processing depends on season. We see a substantial difference between spring and autumn
18 floods using critical success index (CSI) for evaluation. In autumn, there are almost no predictive skill for
19 more than 3 days lead-time. Spring is quite different with a longer prediction horizon; for some catchments
20 and processing schemes the floods are predicted up to nine days in advance. The results indicate a higher
21 predictability in spring floods, which in addition to precipitation are highly dependent on temperature that
22 controls the snowmelt intensity.

23 • The high precipitation rates, which is the flood generating process in autumn, should hence be the focus for
24 further improvements. We found that for some incidents of high precipitation rates the BMA preprocessing
25 resulted in unrealistic precipitation amounts for individual ensemble members. Approaches to amend this are
26 needed.

27 To summarize; we find that flood forecasts benefit from pre- and/or postprocessing, however the optimal processing
28 approaches depend on region, catchment, and season.

## 8  Acknowledgment

## 9  Data and scripts

34 We have used the R-package ncdf4, ensembleMOS, ensembleBMA, SpecsVerification.

1     https://github.com/metno/fimex, was used for the resampling and reprojection of the gridded datasets, and

2     https://github.com/metno/gridpp which includes the preprocessing methods was applied for temperature and

3     precipitation calibration (CAL).

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

5     The ensemble forecast data is available from ECMWF, and streamflow observation is available from NVE upon

6     request

7

## 10 References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. J. Climate, 9, 1518–1530, 1996.

Barnes, L. R., Gruntfest, E. C., Hayden, M. H., Schultz, D. M., Benight, C.: False Alarms and Close Calls: A Conceptual Model for Warning Accuracy, Weather and Forecasting, **2**, 1140-1147, (Corr: 2009, **24**, 1452-1454), 2007

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

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.

Benninga, H.-J. F., Booij, M. J., Romanowicz, R. J., and Rientjes, T. H. M.: Performance of ensemble streamflow forecasts under varied hydrometeorological conditions, Hydrol. Earth Syst. Sci., 21, 5273–5291, https://doi.org/10.5194/hess-21-5273-2017, 2017.

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

Bremnes, J. B.: Improved calibration of precipitation forecasts using ensemble techniques. In Practice, 10, 5, 2007.

Bremnes, J. B.: Improved calibration of precipitation forecasts using ensemble techniques. Part 2: statistical calibration methods, met.no, Report no. 4, 34 pp., Oslo, Norway, available at: http://met-xpprod.customer.enonic.io/publikasjoner/met-report/met-report-2007 (last access: 1 February 2019), 2007.

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

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

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.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M.: Comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Monthly Weather Review, 133(5), 1076-1097, 2005.

1    Clark, M. P., S. Gangopadhyay, L. E. Hay, B. Rajagopalan, and Wilby, R. L.: The Schaake shuffle: A method for

2    reconstructing space–time variability in forecasted precipitation and temperature fields. J. Hydrometeor., 5, 243–262,

3    doi:https://doi.org/10.1175/1525-7541(2004)005<0243, 2004.

4    Cloke, H. L. and Pappenberger, F. Ensemble Forecasting: A review. Journal of Hydrology, 375(3), 613-626, 2009.

5    Ceppi, A., Ravazzani, G., Salandin, A., Rabuffetti, D., Montani, A., Borgonovo, E., and Mancini, M.: Effects of

6    temperature on flood forecasting: analysis of an operative case study in Alpine basins. Natural Hazards and Earth

7    System Sciences, 13(4), 1051., 2013.

8    Constantinou, A., and Fenton, N. E.: Solving the problem of inadequate scoring rules for assessing probabilistic

9    football forecast models, *Journal of the Royal Statistical society*, Series C: Applied Statistics, 2012.

10    ECMWF. Set III - Atmospheric model Ensemble 15-day forecast (ENS). Retrieved from

11    https://www.ecmwf.int/en/forecasts/datasets/set-iii, () 2018a.

12    ECMWF. Changes in ECMWF models. Retrieved from https://www.ecmwf.int/en/forecasts/documentation-and-

13    support/changes-ecmwf-model, 2018b.

14    Engdahl, B.J.K. and Homleid, M.: Verification of experimental and Operational Weather Prediction Models

15    December 2014 to February 2015. Norwegian Meteorological Institute, MetInfo (18/2015), 2015

16    Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the

17    HBV model. Journal of Hydrology, 384(1), 142-155, 2010.

18    Fraley, C., Raftery, A. E., & Gneiting, T.: Calibrating multimodel forecast ensembles with exchangeable and missing

19    members using Bayesian model averaging. Monthly Weather Review, 138(1), 190-202, 2010.

20    Frogner, I-L, Singleton, AT, Køltzow, MØ, Andrae, U. Convection-permitting ensembles: Challenges related to their

21    design and use. Q J R Meteorol Soc. 145 ( Suppl. 1): 90– 106. https://doi.org/10.1002/qj.3525, 2019.

22    Frogner, I., and Coauthors: HarmonEPS—The HARMONIE Ensemble Prediction System. Wea. Forecasting, 34,

23    1909–1937, https://doi.org/10.1175/WAF-D-19-0030.1., 2019

24    Gneiting, T., Raftery, A. E., Westveld III, A.H., and Goldman, T.: Calibrated Probabilistic Forecasting Using

25    Ensemble Model Output Statistics and Minimum CRPS Estimation. Monthly Weather Review, 133(5), 1098-1118,

26    2005.

27    Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness. Journal of the

28    Royal Statistical Society: Series B (Statistical Methodology), 69(2), 243-268. doi:10.1111/j.1467-

29    9868.2007.00587.x, 2007.

30    Gottschalk, L., Jensen, J. L., Lundquist, D., Solantie, R., and Tollan, A.: Hydrologic Regions in the Nordic

31    Countries. Hydrology Research, 10(5), 273-286, 1979.

Hydrology and
Earth System
Sciences
Discussions
Open Access
EGU

1   Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample

2   hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18, 463–477, https://doi.org/10.5194/hess-

3   18-463-2014, 2014

4   Ruan, G.: personal comment 15.06.2016 [Calibration of HBV - NVE flood forcasting], 2016.

5   Hamill, T.M. and Colucci, S. J.: Verification of Eta-RSM Short-Range Ensemble Forecasts. Monthly Weather

6   Review, 125(6), 1312-1327, 1997.

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

8   550-560, 2001.

9   Hanssen-Bauer, I., Førland, E. J., Haddeland, I., Hisdal, H., Mayer, S., Nesje, A., Nilsen, J.E.Ø., Sandven, S., Sandø,

10  A.B., and Sorteberg, A.: Climate in Norway 2100 - a knowledge base for climate adaption. Tech. Rep. 1, Norwegian

11  Climate Service Centre, 2017.

12  Hegdahl, T. J., Engeland, K., Steinsland, I., & Tallaksen, L. M.: Streamflow forecast sensitivity to air temperature

13  forecast calibration for 139 Norwegian catchments. Hydrology and Earth System Sciences, 23(2), 723-739, 2019.

14  Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems.

15  Weather and Forecasting, 15(5), 559-570. doi:10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2, 2000.

16  Jha, S. K., Shrestha, D. L., Stadnyk, T., & Coulibaly, P.: Evaluation of ensemble precipitation forecasts generated

17  through post-processing in a Canadian catchment. Hydrology and Earth System Sciences, 22(3), 1957-1969, 2018.

18  Jolliffe, I. T., & Stephenson, D. B. (Eds.): Forecast verification: a practitioner's guide in atmospheric science. John

19  Wiley & Sons, 2012.

20  Leith, C. E.: Theoretical skill of Monte Carlo forecasts. Monthly weather review, 102(6), 409-418., 1974.

21  Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z.: A review on statistical postprocessing methods for

22  hydrometeorological ensemble forecasting. Wiley Interdisciplinary Reviews: Water, 4(December), e1246.

23  https://doi.org/10.1002/wat2.1246, 2017.

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

25  Madadgar, S., Moradkhani, H., & Garen, D.: Towards improved post-processing of hydrologic forecast ensembles.

26  Hydrological Processes, 28(1), 104-122, 2014.

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

28  Report, 8, 2008.

29  Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I - A discussion of principles.

30  Journal of Hydrology, 10(3), 282-290. doi:10.1016/0022-1694(70)90255-6, 1970.

31  Owens, R. G., and Hewson, T. D.: ECMWF Forecast User Guide. Retrieved from Reading:

32  https://confluence.ecmwf.int/display/FUG/Re-forecasts, doi: 10.21957/m1cs7h, 2018.

Hydrology and
Earth System
Sciences
Discussions
EGU
Open Access

1  Pappenberger, F., Cloke, H. L.,Parker, D. J., Wetterhall, F., Richardson, D. S.,  Thielen, J.: The monetary benefit of

2  early flood warnings in Europe, Environmental Science & Policy, Volume 51, pp. 278-291, doi:

3  10.1016/j.envsci.2015.04.016, 2017.

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

5  Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian model averaging to calibrate

6  forecast ensembles. Monthly Weather Review, 133(5), 1155-1174, 2005.

7  Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting: Uncertainty quantification in complex simulation models using

8  ensemble copula coupling. Stat. Sci., 28, 616–640, doi:https://doi.org/10.1214/13-STS443, 2013.

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

10  Seierstad, I., Kristiansen, J., and Nipen, T.: Better temperature forecasts along the Norwegain coast, newsletter, 148,

11  available at: https://www.ecmwf.int/en/newsletter/148/news/better-temperature-forecasts-along-norwegian-coast

12  (last access: 1 February 2019), 2016.

13  Sheridan, P., Smith, S., Brown, A., and Vosper, S.: A simple height-based correction for temperature downscaling in

14  complex terrain. Meteorological Applications, 17(3), 329-339, 2010.

15  Sloughter, J. Mc Lean, et al. Probabilistic quantitative precipitation forecasting using Bayesian model averaging.

16  Monthly Weather Review, 135.9: 3209-3220, 2007.

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

18  26, 1996.

19  Talagrand, O., R. Vautard, and B. Strauss: Evaluation of probabilistic prediction systems. Proc. ECMWF Workshop

20  on Predictability, Reading, United Kingdom, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading,

21  Berkshire RG2 9AX, United Kingdom.], 1997.

22  Thyer, M., Kuczera, G., & Wang, Q. J.: Quantifying parameter uncertainty in stochastic models using the Box–Cox

23  transformation. Journal of Hydrology, 265(1-4), 246-257, 2002.

24  Tveito, O. E., Bjørdal, I., Skjelvåg, A. O., and Aune, B.: A GIS-based agro-ecological decision system based on

25  gridded climatology, Meteorological Applications., 12, 57–68. https://doi.org/10.1017/S1350482705001490, 2005.

26  Tveito, O. E.: Spatial distribution of winter temperatures in Norway related to topography and large-scale

27  atmospheric circulation, Proceedings of the PUB Kick-off meeting held in Brasilia, 20–22 November 2002. IAHS

28  Publications. 309, 2007, 186-194, 2007.

29  UNISDR: Guidelines for Reducing Flood Losses, United Nations International Strategy for Disaster Reduction,

30  DRR7639 UNISDR, 2004 http://www.unisdr.org/we/inform/publications/558

31  Vannitsem, S., Wilks, D. S., and Messner, J. W., Editor(s): Statistical Postprocessing of Ensemble Forecasts,

32  Elsevier, ISBN 9780128123720, doi: 10.1016/B978-0-12-812372-0.09988-X, 2018.

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

4    Wilks, D. S., & Hamill, T. M.: Comparison of ensemble-MOS methods using GFS reforecasts. Monthly weather
5    review, 135(6), 2379-2390, 2007.

6    Wilson, D., Fleig, A, K,. Lawrence, D., Hisdal, H., Pettersson, L.-E., and Holmqvist, E.: A review of NVE's flood
7    frequency estimation procedures, Norwegian Water Resources and Energy Directorate. Report no. 9-2011, pp 50,
8    (http://publikasjoner.nve.no/report/2011/report2011_09.pdf), 2011.

9    Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R.: Calibrated Surface Temperature Forecast from the
10   Canadian Ensemble Prediction System using Bayesian Model Averaging, Monthly Weather Review, Vol: 134,
11   pp1364-1385, 2007.

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

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

16   Zhu, Y., and R. E. Newell: A proposed algorithm for moisture fluxes from atmospheric rivers. Mon. Wea. Rev., 126,
17   725–735, https://doi.org/10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2., 1998.

18

19

Hydrology and
Earth System
Sciences
Discussions

**11    Tables and Figures**

Table 1 Catchment characteristics for selected catchments: Catchment Area, Annual runoff (Q), Annual precipitation (P),

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 (%) |
|------|------|------|------|------|------|
| Vaekkava | 2078 | 375 | 414 | 0.87 | 0.00 |
| Refsvatn | 53 | 1843 | 297 | 1.00 | 0.00 |
| Tannsvatn | 118 | 719 | 905 | 4.59 | 0.00 |
| Moeska | 121 | 1585 | 325 | 1.71 | 0.00 |
| Nybergsund | 4425 | 487 | 781 | 2.48 | 0.00 |
| Bulken | 1092 | 2038 | 867 | 0.88 | 0.39 |

Table 2 Overview of data and parameters applied the different calibration schemes.

| Variable | Resolution | Reference data | Lead time | Season/ Annual | Training period |
|------|------|------|------|------|------|
| Pcal | Grid ~25km | 200 WMO | - | - | 2014 |
| Tcal | Grid ~25km | Hirlam 5km | Parameters estimated using the first 24 hours, applied to all lead times | Monthly specific parameter values | 2006 to 2011 |
| Pbma | Catchment average | seNorge catchment average | Parameters lead-time specific 1:9 | Parameters specific each issue date | 45 previous days |
| Tbma | Catchment average | seNorge catchment average | Parameters lead-time specific 1:9 | Parameters specific each issue date | 45 previous days |
| Qbma | Catchment average | Sim HBV | Parameters lead-time specific 1:9 | Parameters specific each issue date | 45 previous days |

Table 3 Contingency table for classification of hits (H), missed events (M), false alarms (F), and correct non-events (N).

| | | Observation | |
|------|------|------|------|
| | | **No** | **Yes** |
| **Forecast** | **No** | *N* | *M* |

|  | Yes | *F* | *H* |
|---|---|---|---|

1

2 **Table 4 Summary of the results in Fig 9. ∑Post and ∑Pre shows the number of the catchments where the combination of**

3 **pre- and postprocessing approaches gave the best performance. % pre shows the percentage of catchments where**

4 **preprocessing gave the best performance. Tbma_Praw Traw_Pbam, Tbma_Pbma shows which preprocessing scheme**

5 **using BMA that gave the best performance.**

|  | Pre- or postprocessing | | | Preprocessing – BMA | | | | |
|---|---|---|---|---|---|---|---|---|
| Lead time | ∑Post | ∑Pre | % pre | *Tbma_Praw* | *Traw_Pbam* | *Tbma_Pbma* | *∑bma* | *%bma* |
| 1 | 40 | 40 | 50 | 5 | 5 | 20 | 30 | 75 |
| 5 | 37 | 43 | 54 | 12 | 10 | 13 | 35 | 81 |
| 9 | 31 | 49 | 61 | 22 | 6 | 6 | 34 | 69 |

6

7
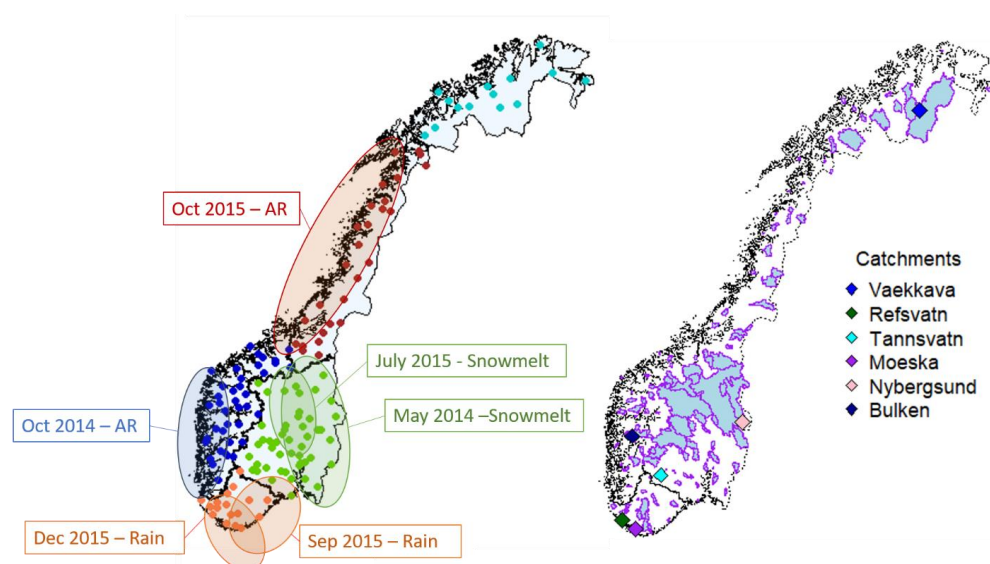


8

9

10 **Figure 1: The map to the left shows the location of the outlet of the 119 catchments used in this study as well as a**

11 **schematic overview of the areas affected by floods caused by different events (rain, snowmelt and atmospheric river (AR))**

12 **during the study period 2014 to 2015. It is worth noting that not all catchments experienced floods within the areas. The**

13 **map to the right shows the catchment areas, and the locations of six catchments for which we will show some detailed**

14 **results are also shown.**

1

2

3

4



5
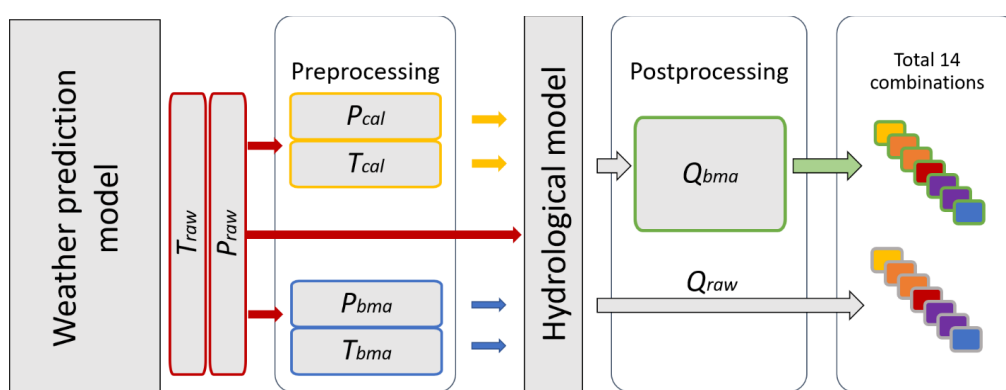
**Figure 2: The processing chain of the experimental set up.** *Traw* and *Praw* are the unprocessed forecasts. Two preprocessing approaches were applied, a grid calibration (CAL) producing the ensembles *Tcal* and *Pcal*, and Bayesian model averaging (BMA) producing the ensembles *Tbma* and *Pbma*. All combinations of *Tcal* and *Pcal* together with *Traw* and *Praw*, as well as all combinations of *Tbma* and *Pbma* together with *Traw* and *Praw*, in total 7 combinations, were run through the hydrological model. BMA was applied to the streamflow forecasts producing the ensembles *Pbma* in addition to *Qraw*. In total 14 combinations of pre- and postprocessing were evaluated. The processing schemes were applied to each issue date, lead time and catchment.
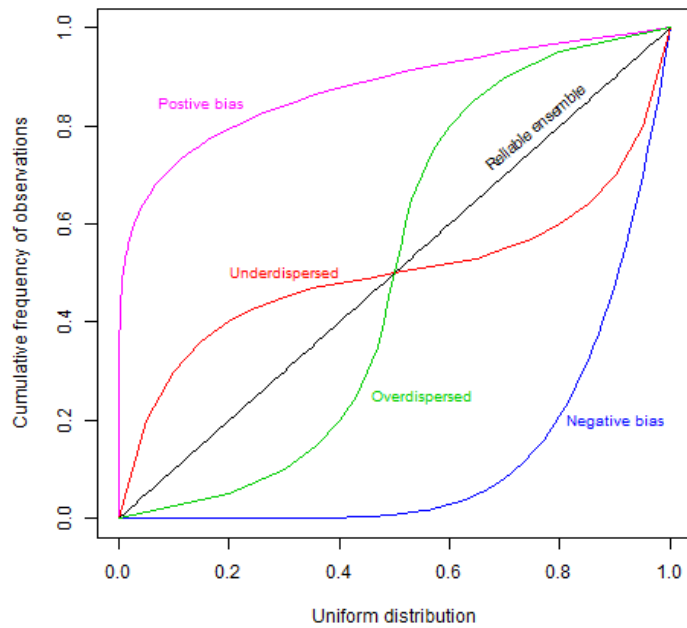
13



14
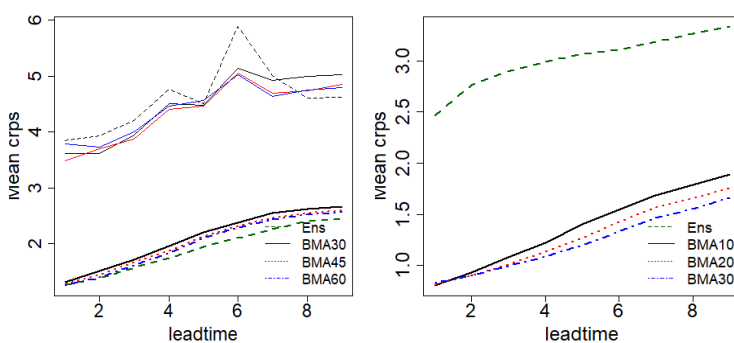
**Figure 3: Left: Precipitation mean CRPS for all lead times for the Aulestad catchment. Thin lines are the 10% percentile precipitation, thicker lines include all the data. Right: temperature mean CRPS for all lead times for Viksvatn.**

17

1

**Figure 4: Typical shapes of the cumulative rank-histogram plots that can be used to detect both biased, over- and underdispersed ensembles. The closer the curves are to the 1:1 line, the more reliable are the ensembles.**

29
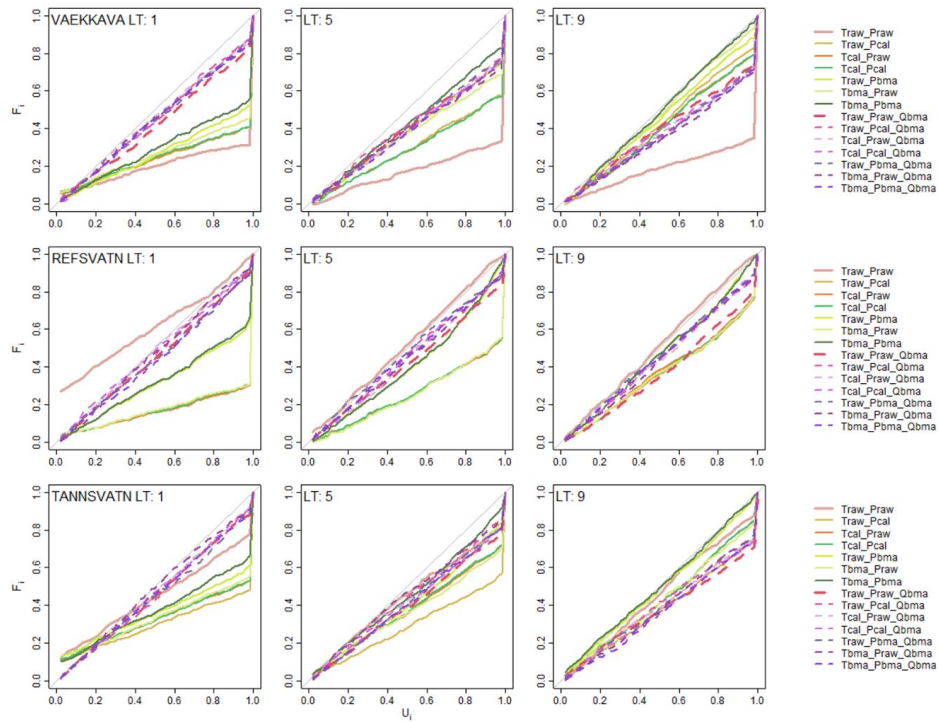
**Figure 5: Reliability plots that compare all 14 processing schemes for Lead time 1, 5 and 9 days (LT: 1,5,9))for three catchments. The location of the catchments is shown in Fig 1 right. The cumulative empirical rank-histograms scaled to unity is shown on the y-axis whereas the uniform distribution is shown on the x-axis. The most reliable forecasts are closest to the 1:1 line. Fig. 4. provides details for interpretation of these plots.**
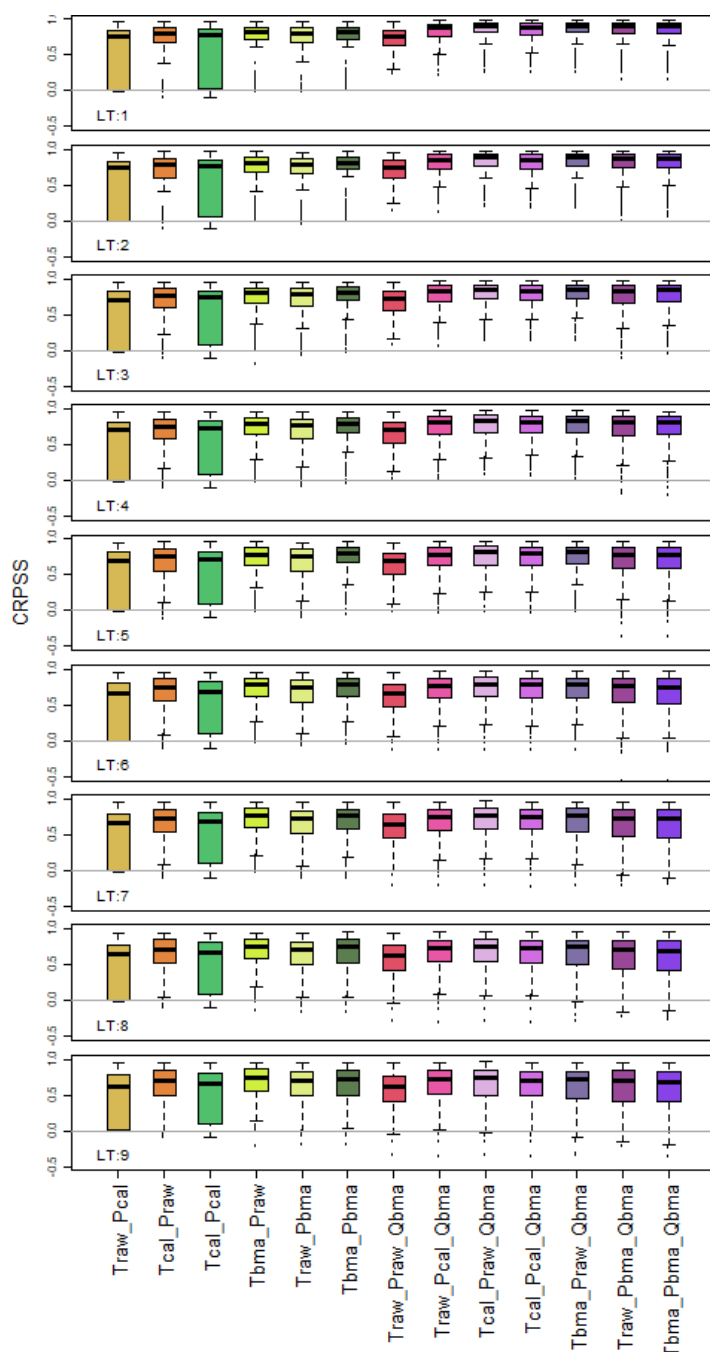
Figure 6: Boxplot of CRPSS (best is 1) for all catchment based on the <u>full dataset</u> for all processing schemes (x-axis) and all lead times (rows). The first six boxplots indicate the different preprocessing schemes, whereas the last seven indicates processing schemes that includes a postprocessing step.

Hydrology and
Earth System
Sciences
Discussions

Open Access

EGU
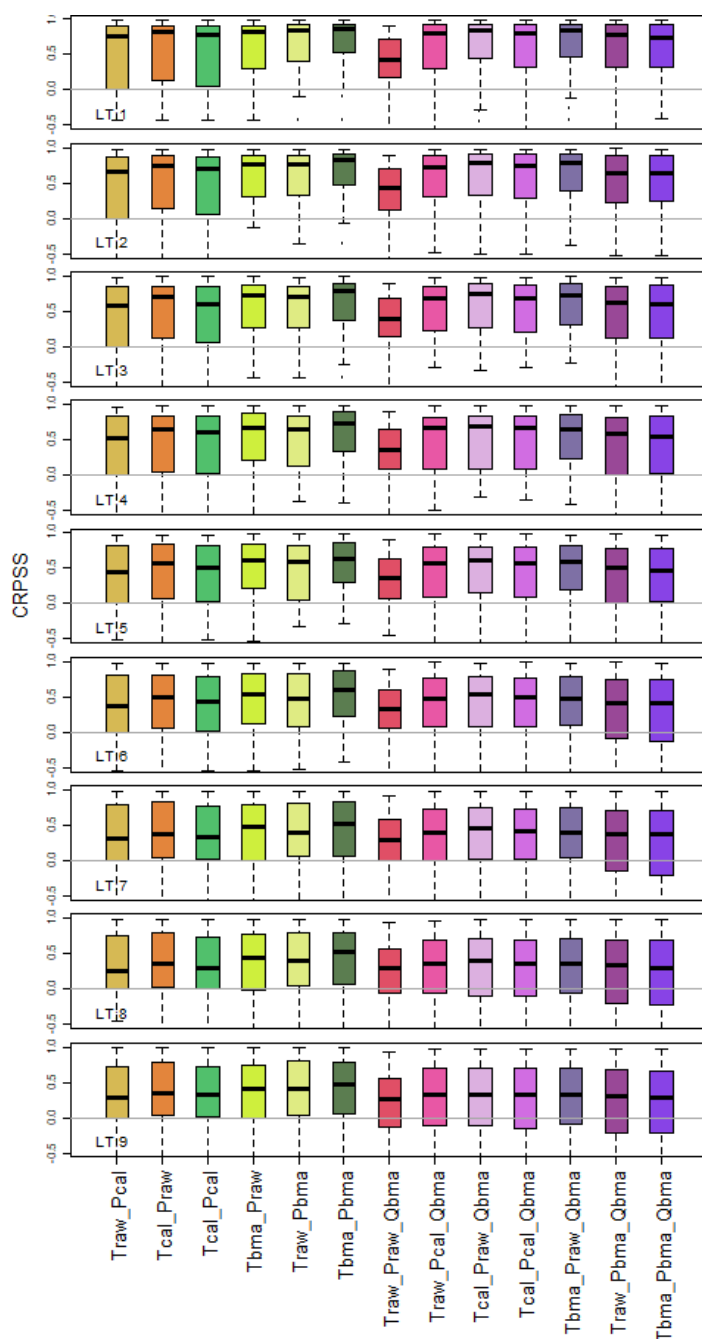
**Figure 7: Boxplot of CRPSS (best is 1) for all catchment based on the <u>flood event dataset</u> for all processing schemes (x-axis) and all lead times (rows). The first six boxplots indicate the different preprocessing schemes, whereas the last seven indicates processing schemes that includes a postprocessing step.**
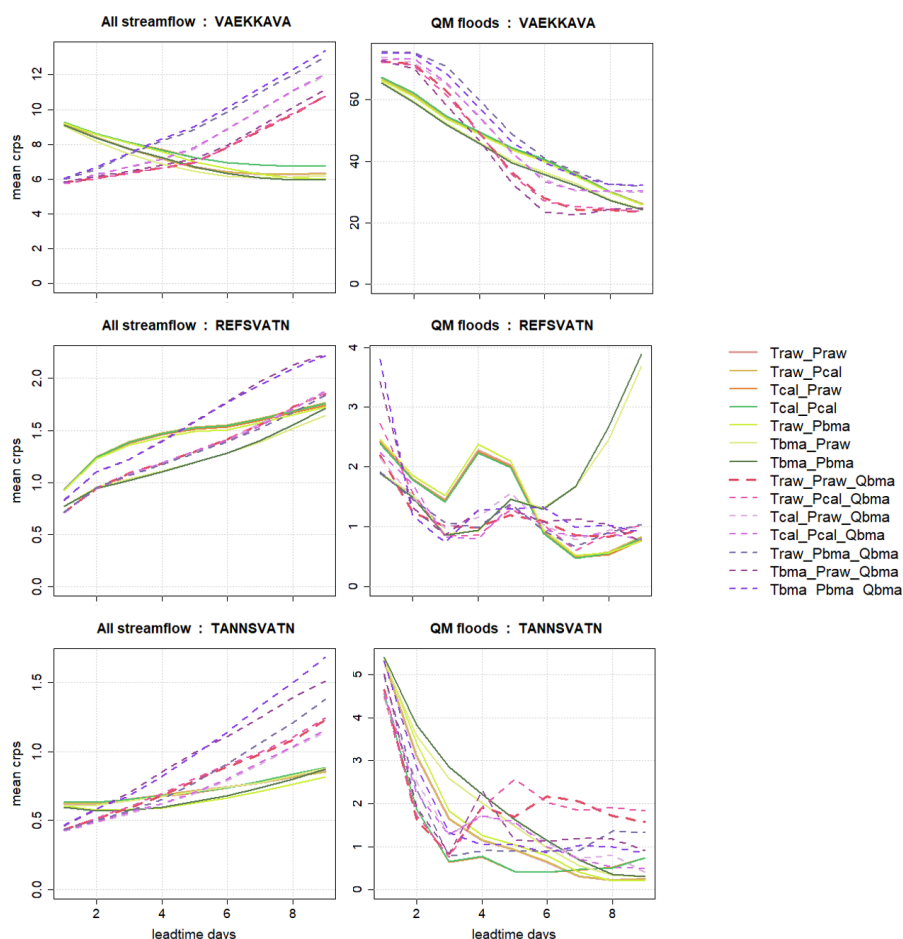
**Figure 8: Mean CRPS (in m³s⁻¹) for three selected catchments as a function of lead time calculated for the full dataset to the left, and the flood dataset to the right. Note that the values for "mean CRPS" on the y-axis is different for the different plots. For Vaekkava 13 days used to calculate the flood dataset (May 30 - June 5 2014 and May 24-30 2015), whereas 2 days were used for Refsvatn (December 05-06 2015) and Tannsvatn (May 21-22 2014).**
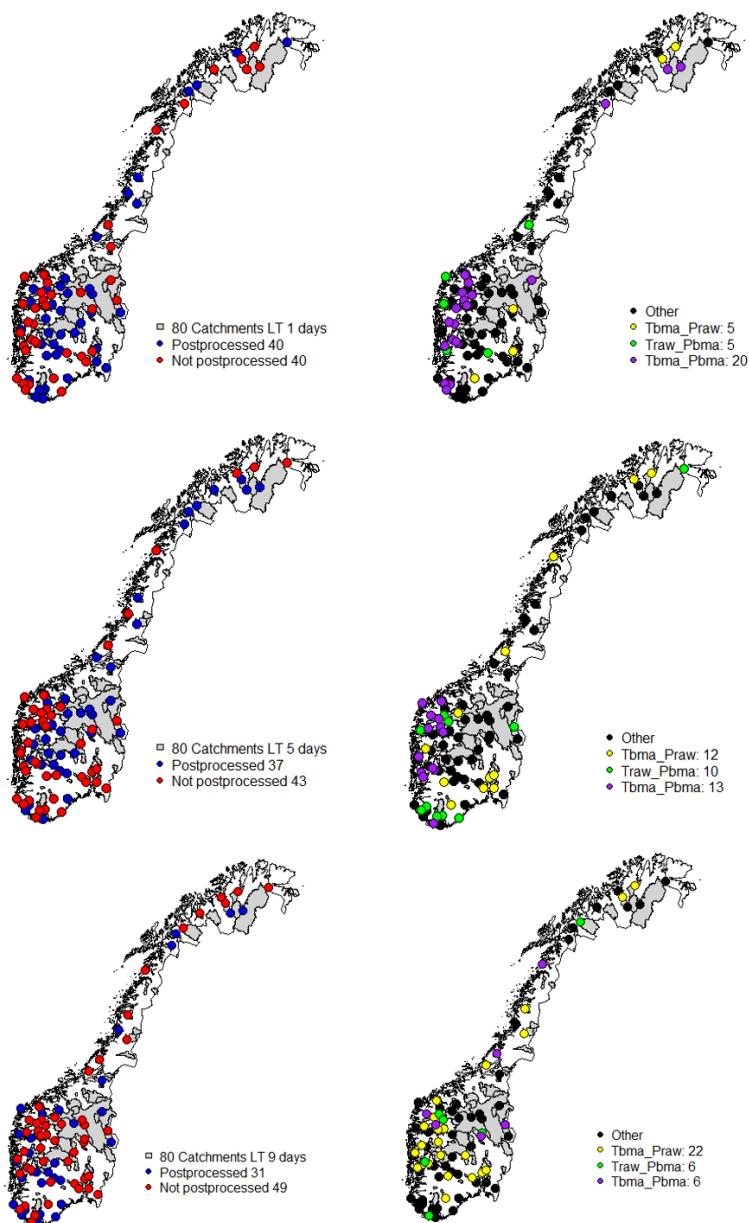
1

2 **Figure 9: Figures to the left indicates catchments where any preprocessing approaches alone (red dots) or the combination**

3 **of pre- and postprocessing (blue dots) provides the highest performance evaluated by the mean CRPS for lead times of 1,**

4 **5, and 9 days. The figures to the right show the BMA preprocessing scheme that provides the best CRPS. All evaluation of**

5 **CRPS was applied for the subset of floods.**

1
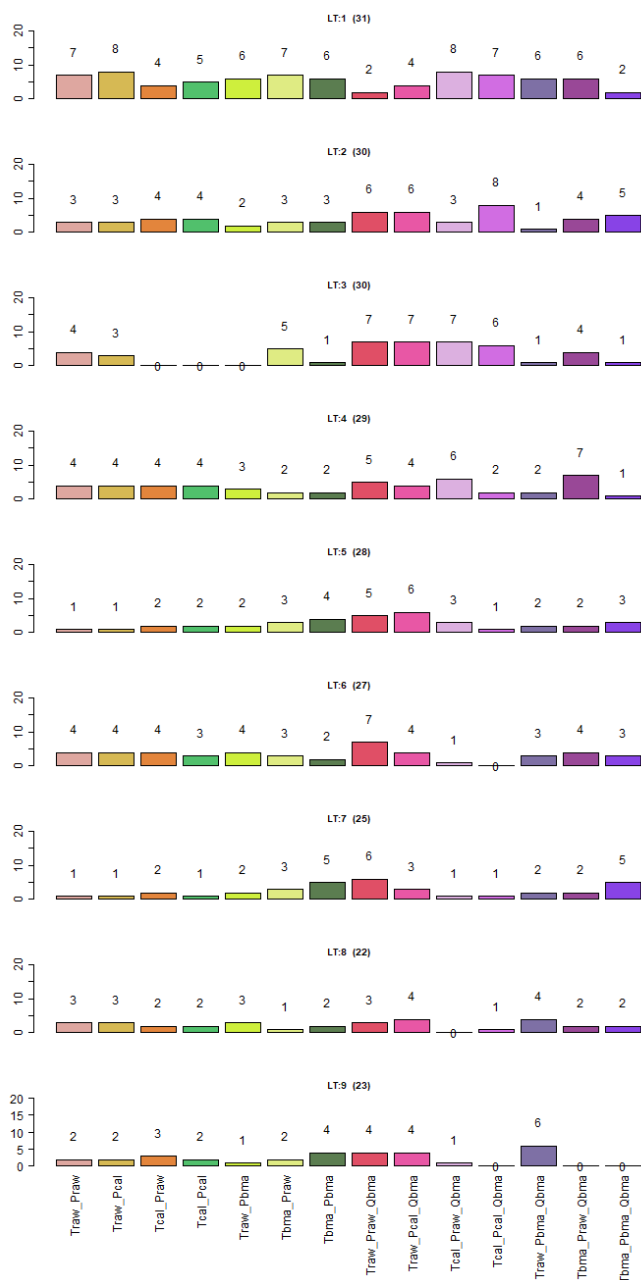
**Figure 10: Spring- Critical success index (CSI). Each row represents one lead time (from 1 to 9 days) and includes all**

**processing schemes. In parenthesis the total number of catchments that predicted the exceedance of warning level.**
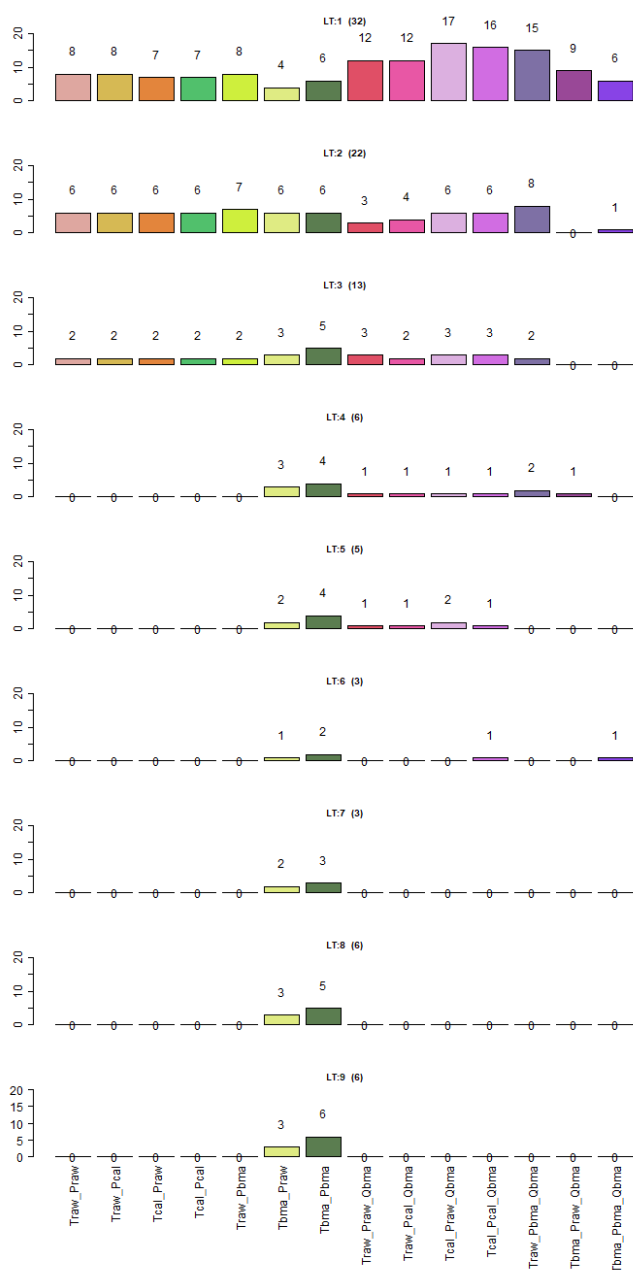
Figure 11: Autumn- Critical success index (CSI). Each row of barplots represent one lead time (from 1 to 9 days) and includes all processing schemes. In parenthesis the total number of catchments that predicted the exceedance of warning level.
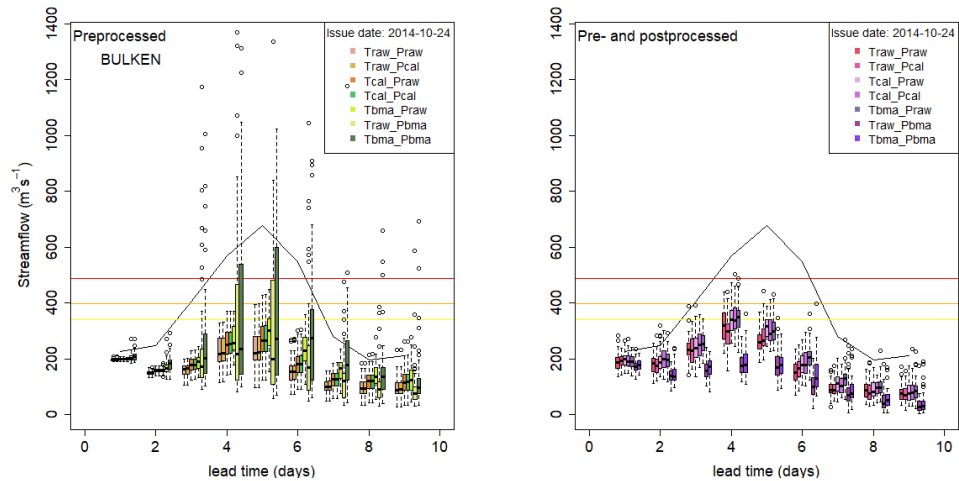
**Figure 12: The AR event 2014 at Bulken. Boxplots of the applied processing schemes. The black line indicates the reference streamflow for the event. The horizontal lines represent the mean annual flood (yellow), the 5-year flood (orange)and the 50-year flood (red).**
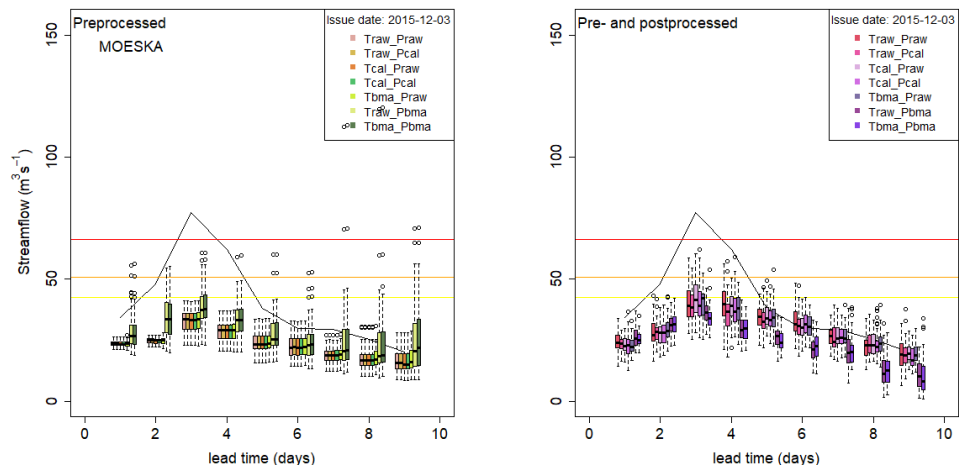


**Figure 13:The extreme weather event Synne in 2015 at Moeska with boxplots indicating the streamflow estimates for different processing approaches. Reference streamflow for the event is the black line. The horizontal lines represent the mean annual flood (yellow), the 5-year flood (orange)and the 50-year flood (red).**
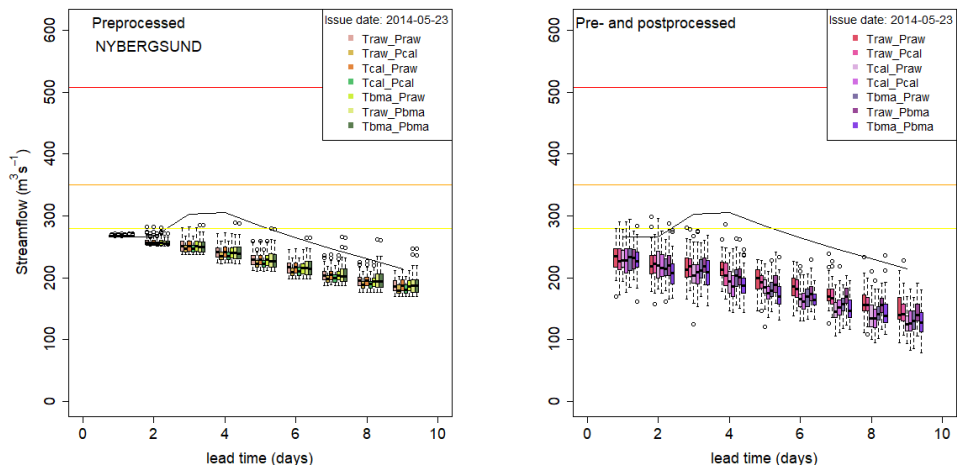
Figure 14: The snowmelt flood in May 2014 at Nybergsund, with boxplots indicating the streamflow estimates for different processing approaches. Reference streamflow for the event is the black line. The horizontal lines represent the mean annual flood (yellow), the 5-year flood (orange)and the 50-year flood (red).

## 11.1 Appendix-Figures
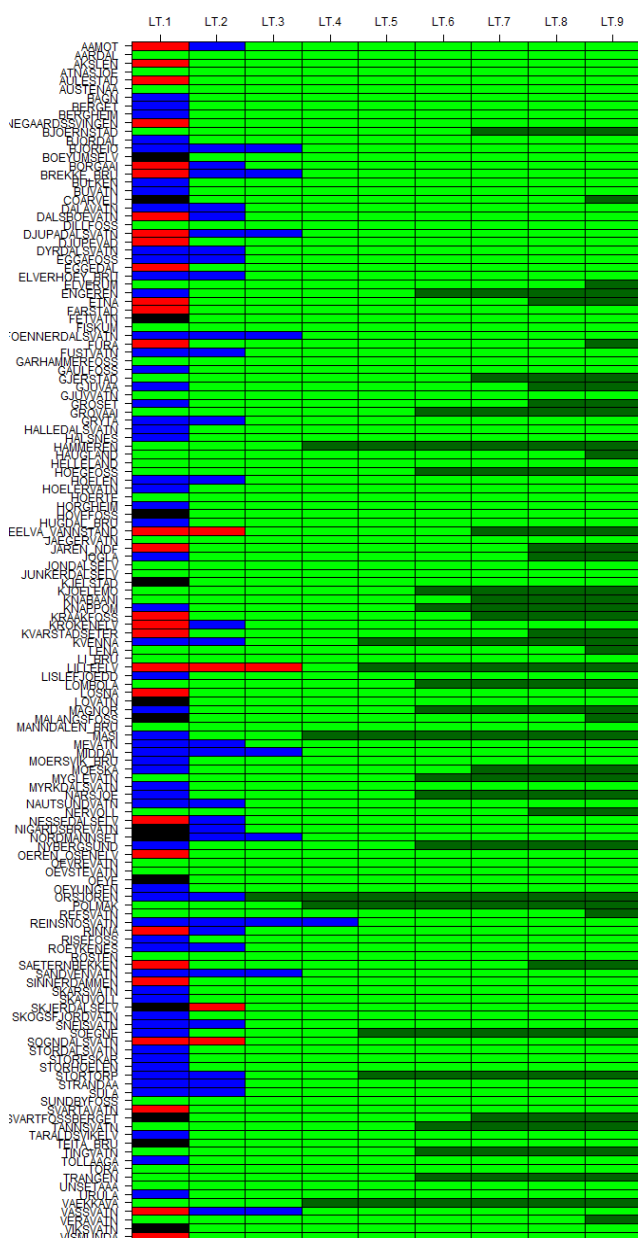


A-Figure 1: Optimal training length for temperature forecasts, using CRPS as evaluation criterion for all catchments (rows) and all lead-times (columns). Table indicates all catchments (rows) and lead-times (columns). Black: 10 days, red: 20 days, blue: 30 days, green: 45 days, dark green: raw ensemble.

Hydrology and
Earth System
Sciences
Open Access
Discussions
EGU

1

2   **A-Figure 2: Optimal training length for precipitation forecasts, using CRPS as evaluation criterion for all catchments**

3   **(rows) and lead-times (columns). Red: 20 days, blue: 30 days, green: 45 days. BMA applied to precipitation depends on**

4   **sufficient number of precipitation values above 0 to converge. We found that for some catchments this was a problem, and**

5   **this was also important for the decision to use a 45 days training window, even though the results from the figure shows**

6   **that for some catchments and lead times, the CRPS is better for shorter training lengths.**

7

1

2  A-Figure 3: All data used to evaluate the best CRPS achieved by applied processing schemes, shown for all catchments

3  and lead times. The color in each cell represent the processing scheme with the best CRPS score. Summary of the results

4  shown in A-Figure 4.
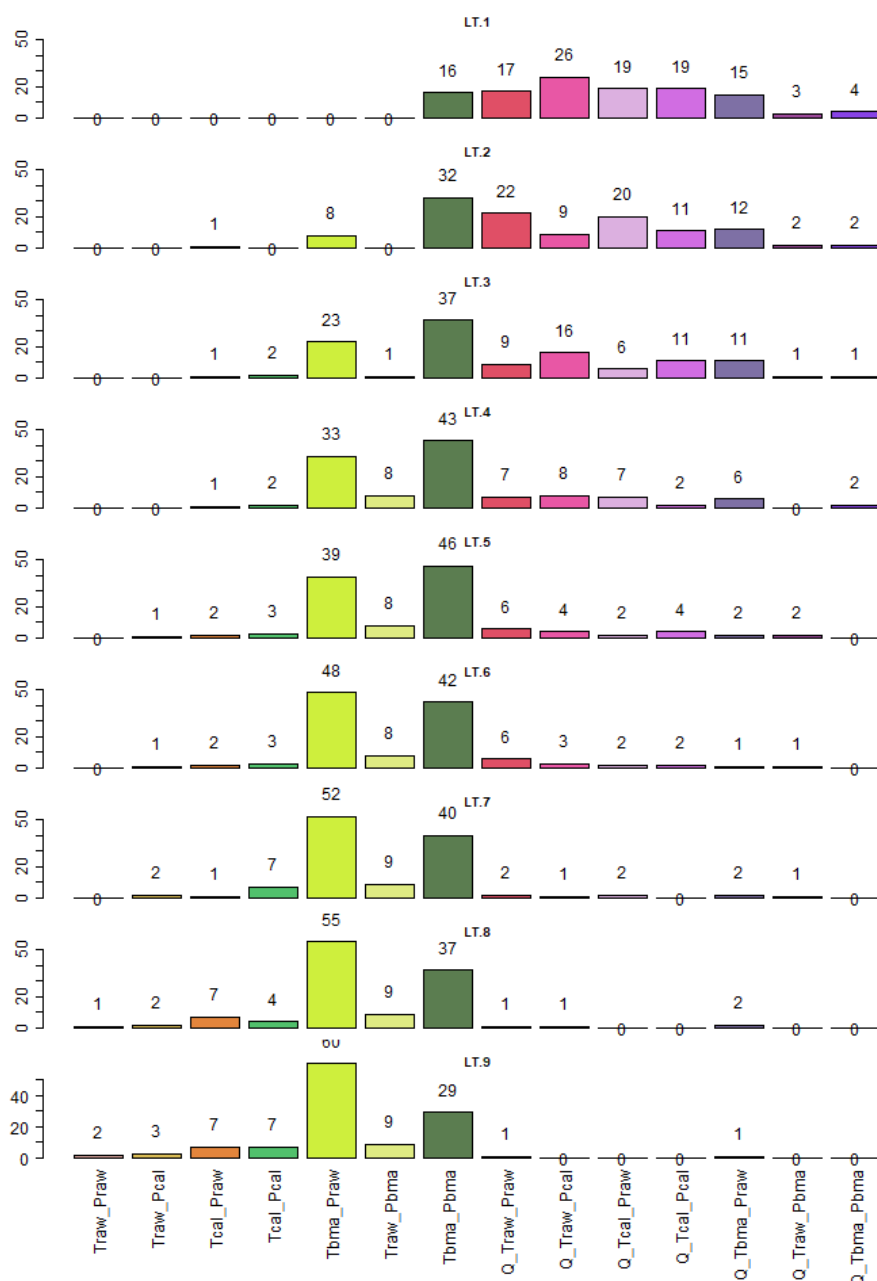
1

2     **A-Figure 4: Summary of Figure 3. Evaluation applied to all data. Each bar indicates the number of catchments for which**

3     **the specific processing scheme attained the best CRPS score. Lead-times from 1 day (top) to 9 days (bottom).**

**A-Figure 5: Flood dataset used to evaluate the best CRPS achieved by applied processing schemes, shown for all catchments and lead times. The color in each cell represent the processing scheme with the best CRPS score. Summary of the results shown in A-Figure 6.**
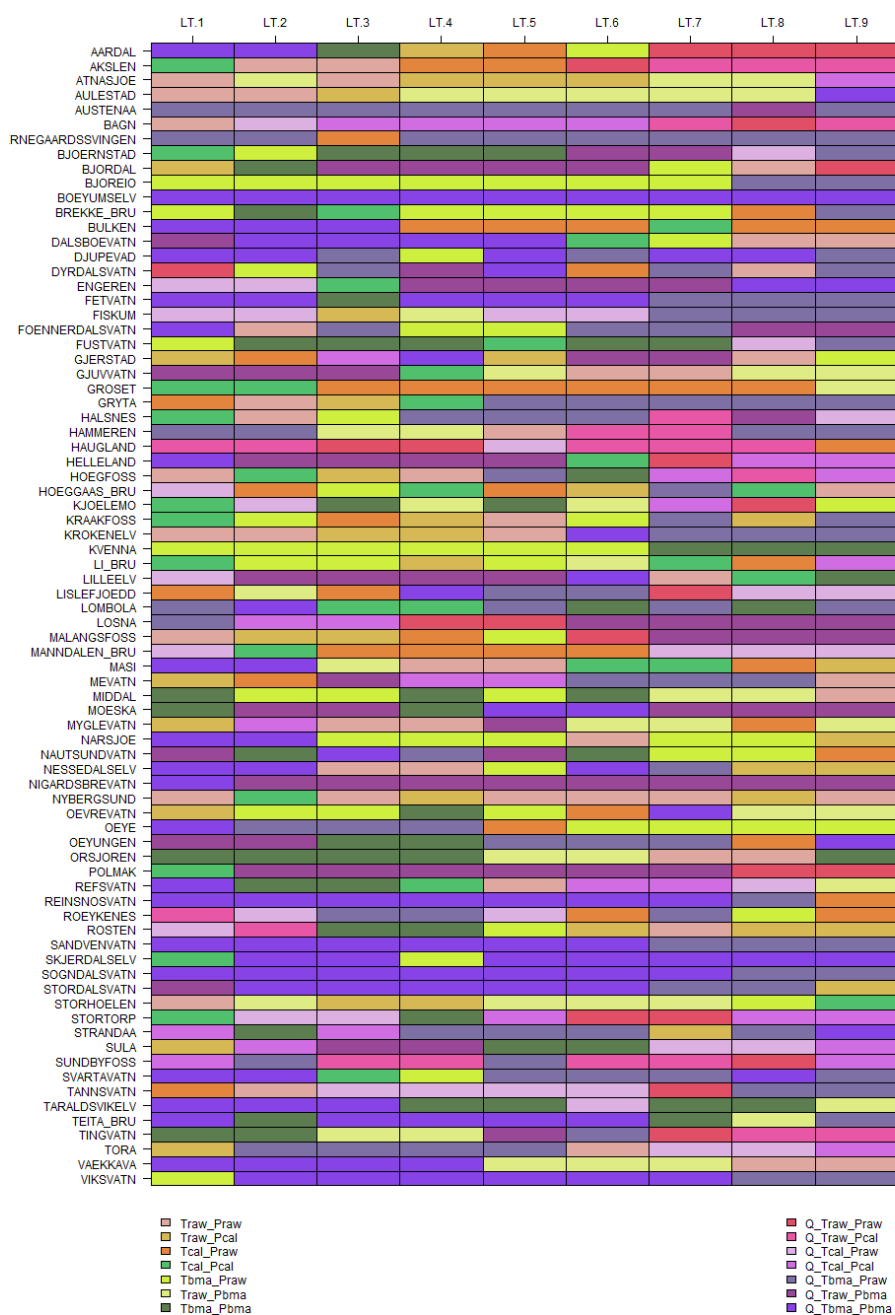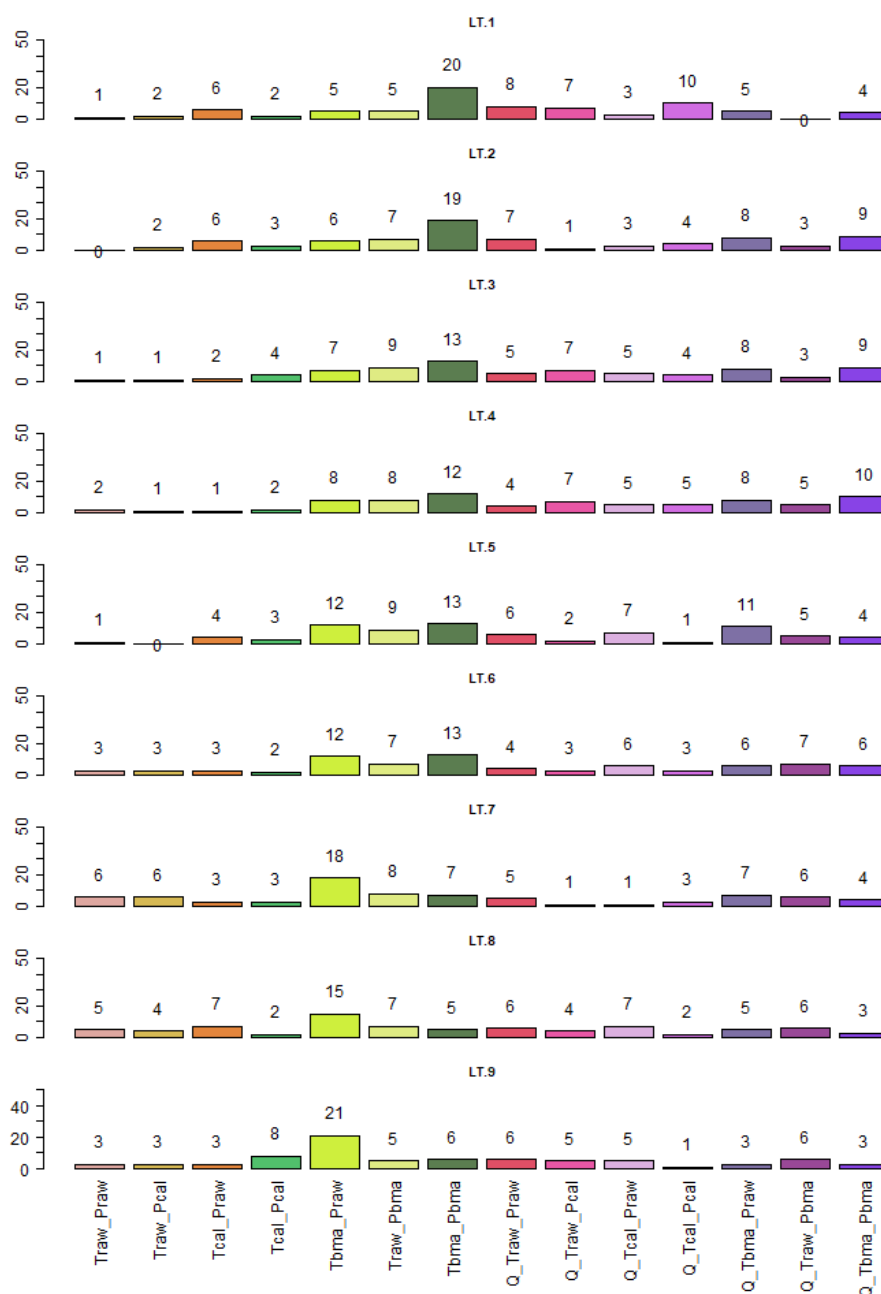
**A-Figure 6: Summary of Figure 5. Evaluation applied to the flood dataset. Each bar indicates the number of catchments for which the specific processing scheme attained the best CRPS score. Lead-times from 1 day (top) to 9 days (bottom).**
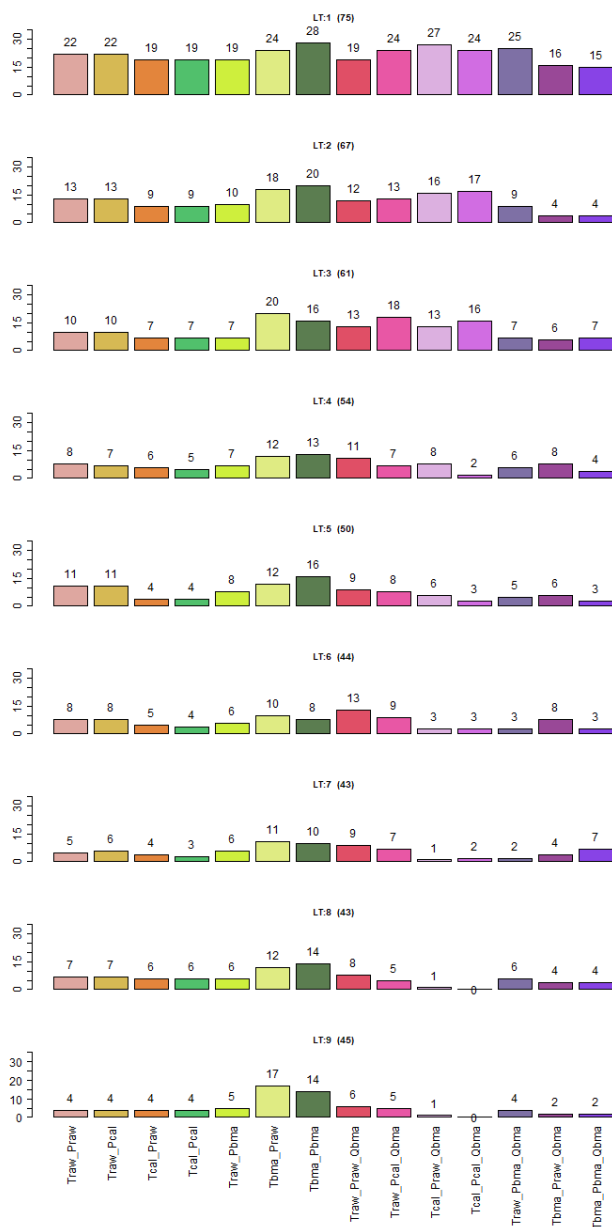
A-Figure 7: CSI for all catchment (86) where there was either forecasted or observed floods during the 2-year period of the study. In this figure, multiple methods can achieve the criteria for exceeding the flood warning level. The number in parenthesis shows the number of catchments where one or more methods successfully indicates the warning level is exceeded.

Paper III:

# An Event-Based Approach to Explore Selected Present and Future Atmospheric River–Induced Floods in Western Norway

**T. J. Hegdahl, K. Engeland, M. Müller, and J. Sillmann**

III

# An Event-Based Approach to Explore Selected Present and Future Atmospheric River–Induced Floods in Western Norway

TRINE J. HEGDAHL

*Norwegian Water Resources and Energy Directorate, Oslo, Norway*

KOLBJØRN ENGELAND

*Norwegian Water Resources and Energy Directorate, and University of Oslo, Oslo, Norway*

MALTE MÜLLER

*Norwegian Meteorological Institute, Oslo, Norway*

JANA SILLMANN

*Center for International Climate Research, Oslo, Norway*

## ABSTRACT

The aim of this study is to investigate extreme precipitation events caused by atmospheric rivers and compare their flood impact in a warmer climate to current climate using an event-based storyline approach. The study was set up by selecting four high-precipitation events from 30 years of present and future climate simulations of the high-resolution global climate model EC-Earth. The two most extreme precipitation events within the selection area for the present and future climate were identified, and EC-Earth was rerun creating 10 perturbed realizations for each event. All realizations were further downscaled with the regional weather prediction model, AROME-MetCoOp. The events were thereafter used as input to the operational Norwegian flood-forecasting model for 37 selected catchments in western Norway, and the magnitude and the spatial pattern of floods were analyzed. The role of the hydrological initial conditions, which are important for the total flooding, were analyzed with a special emphasis on snow and soil moisture excess. The results show that the selected future extreme precipitation events affected more catchments with larger floods, compared to the events from present climate. In addition, multiple realizations of the meteorological forcing and four different hydrological initial conditions, for example, soil saturation and snow storage, were important for the estimation of the maximum flood level. The meteorological forcing (e.g., the internal variability/perturbed output) accounts for the highest contribution to the spread in flood magnitude; however, for some events and catchments the hydrological initial conditions affected the magnitudes of floods more than the meteorological forcing.

## 1. Introduction

Atmospheric rivers are transient, narrow routes of water vapor supplying a substantial fraction of the moisture transport from tropical or extratropical latitudes toward the poles (Zhu and Newell 1998; Ralph and Dettinger 2011; Ralph et al. 2017). When such air masses with a high moisture content reach a topographical barrier like the west coast of Norway, the air parcels are lifted and adiabatically cooled, forming clouds and precipitation (Stohl et al. 2008). For western Norway, the most extreme precipitation, flood, and landslide events since 1900 can largely be attributed to atmospheric rivers (Stohl et al. 2008; Lavers and Villarini 2013, 2015; Azad and Sorteberg 2017; Benedict et al. 2019). Three recent examples of atmospheric rivers unfolded and affected western Norway quite differently. In September 2005, an atmospheric river hit the city of Bergen, with a precipitation intensity that was record

high, and measured to 156.5 mm $(24 h)^{-1}$ and 111 mm $(12 h)^{-1}$ (Iden et al. 2005; Stohl et al. 2008). The precipitation intensity triggered landslides and floods causing locally large damages and casualties. In October 2014, high precipitation of long duration, which coincided with initially high groundwater and soil moisture content, caused large floods in multiple catchments (Langsholt et al. 2015). In October 2018, a cold period allowing snow to accumulate at high altitudes was followed by a very warm atmospheric river event. The combination of rainfall and extensive snowmelt caused floods and large damages, affecting areas not usually susceptible to autumn floods (Vannforeningen 2018). These examples of atmospheric river induced floods demonstrate that a combination of several factors contribute to the total flood impact: the nature of the atmospheric river event itself, including moisture content, landfall, dynamic development, precipitation intensity and duration, as well as the antecedent weather that defines the hydrological initial conditions. Due to the important contribution from snowmelt and the initial soil moisture, there is no unique relationship between precipitation intensities or volumes and flood sizes (Berghuijs et al. 2019).

In the present climate, the Norwegian west coast is one of the wettest parts of Europe with annual precipitation of more than 3000 mm (Stohl et al. 2008). Rain is the major contributor to floods, whereas the contribution from snowmelt increases with elevation and distance from the coast (Kobierska et al. 2018). The majority of and the largest flood events in this region occur in the autumn or early winter (Roald 2008), which overlaps with the main seasons for atmospheric rivers (Azad and Sorteberg 2017). Climate projections for western Norway indicate increased precipitation in the future, where both the number of days with intense rainfall and the intensity of the rainfall will increase (Caroletti and Barstad 2010; Hanssen-Bauer et al. 2017). Baatsen et al. (2015) show that the changes in diabatic heating and moisture transport due to a warmer Atlantic Ocean will cause more severe storms over western Europe and affect storm paths. Storms hitting the British Isles today might move more toward Scandinavia in the future.

To investigate future climate extreme precipitation and floods, often an ensemble of climate models is used to obtain probabilities for future extremes (e.g., Sillmann et al. 2013; Hanssen-Bauer et al. 2017). The standard approach is to apply a climate–hydrological modeling chain that includes an ensemble of global climate models (GCMs), regional climate models (RCMs), and/or statistical downscaling methods, and hydrological models (Olsson et al. 2016). For instance, Lawrence and Hisdal (2011) and Lawrence (2016) estimate the flood probabilities and frequencies for Norwegian rivers based on continuous simulations from such ensembles. Several studies show that the floods in western Norway will increase due to increased future precipitation, and shift toward more rain-induced floods in autumn and winter (e.g., Sorteberg et al. 2018; Hanssen-Bauer et al. 2017; Vormoor et al. 2016; Lawrence and Hisdal 2011).

For Norway, the information extracted from the climate projection studies guides the societal adaptation strategies. The projection studies are the basis for tailored guidelines for climate adaptation for communities (http://www.klimatilpasning.no/infosider/english/), provided by the Norwegian Center for Climate Services (https://klimaservicesenter.no/). One example of a current practice is the use of flood inundation maps that include estimated future flood levels for specific return periods, based on the expected change in streamflow (e.g., Orvedal and Peereboom 2014; https://gis3.nve.no/link/?link=flomsone). The flood inundation maps indicate the water level during a 200-yr flood in a future climate, and these maps are hence the basis for land-use planning and govern the placement of buildings and important infrastructure to avoid future flooding. In addition, the future climate is important for the hydropower industry in Norway, both in terms of dam safety and water available for electricity production.

For most purposes, a multimodel GCM modeling approach is favored. There are, however, instances where one or few models are preferable (IPCC 2010). Some GCMs might resolve and describe specific weather processes better than other GCMs. An example of a weather process that is better described by higher-resolution GCM is the landfall of atmospheric rivers causing orographic precipitation over western Norway. For atmospheric rivers, the model description of the topographical barrier is of the utmost importance to get well-represented precipitation by mountains (Neiman et al. 2009). Large errors might arise in a steep terrain where the elevation and hence precipitation varies greatly, especially for small catchments where the area is smaller than the grid resolution of a GCM. In a future climate, more atmospheric rivers will make landfall while temperatures are above the freezing point, and thereby deposit less snow and more rain due to higher mean temperature (Whan et al. 2020) and hence influence the seasonality of atmospheric river induced floods. While atmospheric rivers and their future changes have garnered recent attention in climate studies (e.g., Dettinger 2011; Ralph and Dettinger 2011; Espinoza et al. 2018), their ultimate effect on catchment-level flows is not as well studied.
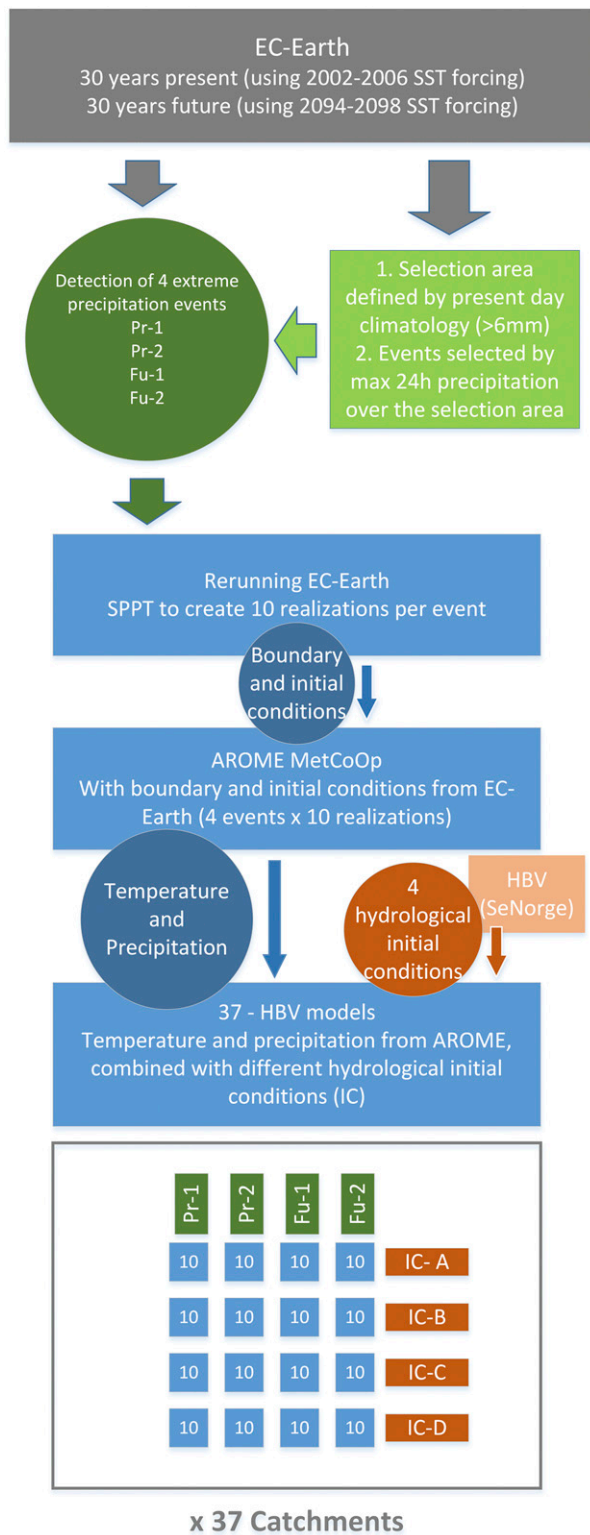
FIG. 1. The processing chain starting from EC-Earth climate projections, using SST forcing (Hazeleger et al. 2010; Haarsma et al. 2013). Further describing the event selections where the maximum 24-h precipitation over western Norway is chosen. For each event, stochastic perturbation of physical tendencies (SPPT)

The aim of this study is to analyze the impact of extreme atmospheric river events in western Norway under present and future climate by using an event-based storyline approach as outlined in Hazeleger et al. (2015). Shepherd et al. (2018) defines a storyline as "a physical self-consistent unfolding of past events, or of possible future events or pathways." Moreover, Shepherd et al. (2018) argue that one reason for applying a storyline approach is to improve awareness of risk based on plausibility rather than probability. The storyline approach is particularly useful for decision makers since it enables them to assess "impact of particular actions under an uncertain regional climate change" (Shepherd 2019). By applying an event-based storyline approach and selecting only a few events, it is easier to do more computer-intensive high-resolution global and regional modeling than is possible with a coarse resolution multi-model ensemble of GCMs. High-resolution climate modeling is important for our study area with complex topography that is not captured in coarse resolution GCMs (Prein et al. 2015). The latter will often result in large precipitation and/or temperature biases that would need correction before being used as input for, for example, hydrological impact modeling (e.g., Maraun 2016). In this study, we used a modeling chain similar to the operational flood-forecasting modeling chain, and thereby familiar to stakeholders in Norway. By using a familiar modeling chain, established exceedance thresholds, and warning colors, it is easier to assess implications of future floods and it facilitates the communication about future flood impact and risk awareness. The comprehension and utilization of climate change data depend on the user (Porter and Dessai 2017; Howarth et al. 2017), and there is often a mismatch between the scientist perception of what the user need and what the user wants. We believe that the approach used in this study is a contribution to mitigating this mismatch.

The event-based storyline approach used for this study comprises the entire modeling chain from a high-resolution global climate model, to a regional weather prediction model and finally a hydrological model estimating floods (Fig. 1). A novel aspect is that the two

is used to establish new event realizations by rerunning EC-Earth. All EC-Earth realizations are downscaled using the regional weather forecasting model AROME-MetCoOp. Temperature and precipitation data from AROME-MetCoOp are the input to the hydrological models (HBV), which are run for all selected events in combination with four different hydrological initial conditions. The hydrological initial conditions are determined by running HBV with seNorge-interpolated observational data.

downstream models are similar to the operational weather and flood-forecasting models, and thus optimally calibrated to the region of interest. Further, the global climate model has a spatial resolution comparable to that of the global weather prediction model used operationally. With this model setup, we investigated atmospheric river driven extreme precipitation events for western Norway under present and future climates. In addition to the change in precipitation magnitude and intensity in a future climate, the hydrological initial conditions might also change, and should be addressed more specifically (e.g., Sharma et al. 2018). We therefore selected a set of four different characteristic initial conditions that we combined with all atmospheric river events. The hydrological initial conditions were established using historical data from selected years to spin up the hydrological model. This approach enabled us to consider the importance of the hydrological initial states when we wanted to represent different future flood scenarios. Furthermore, we highlighted different flood responses to the atmospheric river events and evaluated the relative importance of hydrological initial conditions and multiple meteorological model realizations on the flood estimations.

## 2. Models and methods

### a. Global climate model

We used simulations from the global climate model EC-Earth v2.3 (Hazeleger et al. 2010; Haarsma et al. 2013) with a resolution of about 25 km (T799L91). High-resolution global climate models with spatial resolutions of around 30 km are better capable of simulating the water transport over the Atlantic compared to coarse resolution global models, and hence can better represent small-scale extreme weather systems (Haarsma et al. 2013). Two different model ensemble simulations were considered, a present-day scenario from 2002 to 2006 and a future scenario from 2094 to 2098. Each scenario consists of an ensemble of six independent members, for which the initial atmospheric conditions are determined by running a low-resolution model (EC-Earth, T159) 10 years, and use one of the first six days after the spinup period as atmospheric conditions for each of the six ensemble members. The EC-Earth model (T799L91) is run for 3 months, to ensure that the members are independent. Thereafter, the six independent members are run for 5 years starting 1 January, which for each period results in a 30-yr dataset. In the present-day simulations, observed greenhouse gas and aerosol concentrations are applied, while in the future simulation the concentrations are derived from the representative concentration pathway (RCP) 4.5 scenario (van Vuuren et al. 2011).

The sea surface temperatures (SSTs) are used as lower boundary condition. For the period from 2002 to 2006, a daily SST satellite product (http://www.ncdc.noaa.gov/oa/climate/research/sst/oi-daily.php) is used. For the period from 2094 to 2098, the ensemble mean changes in SST from future projections of a coupled atmosphere–ocean climate model using the SRES A1B scenario (ESSENCE project; Sterl et al. 2008; Haarsma et al. 2013) is added to the 2002–06 SST. The projected global temperature change at the end of the century under SRES A1B lies within the CMIP5-projected range under RCP4.5 (Sillmann et al. 2013).

Haarsma et al. (2013) provide a detailed description of the EC-Earth model setup. The EC-Earth simulations are previously used and validated in several studies (e.g., van der Linden et al. 2018; van Haren et al. 2015; Bintanja et al. 2014; Baatsen et al. 2015; Haarsma et al. 2013). Further, Whan et al. (2020) show that the present EC-Earth simulations, when compared to ERA-Interim (Dee et al. 2011), are able to represent both the frequency and intensity of atmospheric river precipitation in present climate. The integrated water vapor transport (IVT) is a measure used to track and define the atmospheric rivers. Whan et al. (2020) used an automatic algorithm to track the atmospheric rivers from 6-hourly integrated water vapor transport and defined the IVT exceedance threshold for the present climate period to $368 \, \mathrm{kg \, m^{-1} \, s^{-1}}$ (95% percentile). In Fig. 2, we find that the 98% percentile of the IVT in EC-Earth is similar to ERA-Interim, which confirms that results from EC-Earth can be used to identify atmospheric rivers.

### b. Event selections

Since nearly all large-scale precipitation extremes at the west coast of Norway are connected to atmospheric rivers (Lavers and Villarini 2015; Azad and Sorteberg 2017; Benedict et al. 2019), the events were selected by identifying the two largest daily precipitation values in a predefined area. The spatial pattern of precipitation is to a large degree controlled by orographic mechanisms. Therefore, the area used to select the events was defined by grid cells where the EC-Earth simulated extreme precipitation, constrained within 57.1°–63.2°N and 2.6°–9.3°E, derived from the 30-yr present-day climatology, exceeded an average of 6 mm day$^{-1}$. The grid cells used in the selection of events for western Norway are shown as a gray shaded area in Fig. 3. The two most extreme precipitation events within the selection area for the present and future climates were identified. From a 30-yr daily dataset, the two highest daily values represent the 99.98% percentile. The two present climate EC-Earth events had a daily average precipitation over the selection area of 84 and 74 mm day$^{-1}$, whereas the

FIG. 3. The gray shaded area defines grid cells in the EC-Earth model used to select the atmospheric river events over the west coast of Norway, whereas the location and catchment area of the 37 catchments are marked in blue with a pink outline. The red box in Fig. 4 defines the extent of this map.

filaments of IVT originating in the Atlantic (Fig. 4). For each selected event, 10 alternative realizations were generated in EC-Earth by stochastically perturbing the model physics tendencies (SPPT) 5 days prior to the selected event in EC-Earth. The SPPT method is similar to that used for the operational ensemble forecasting at the European Centre for Medium-Range Weather Forecasts (ECMWF; Persson 2015).

### c. Regional model

To obtain more realistic values of extreme precipitation, AROME-MetCoOp a nonhydrostatic weather forecasting system (Müller et al. 2017), was utilized for downscaling the $4 \times 10$ realizations from EC-Earth. AROME-MetCoOp is the operational weather forecasting system for Norway, Sweden and Finland (domain defined in Fig. 4) and is used as input for the operational hydrological forecasting systems in Norway. The AROME-MetCoOp model has a spatial resolution of 2.5 km and is initialized and forced at the lateral boundaries by ECMWF IFS in the operational setup, which was replaced by the EC-Earth realizations for this model setup. The simulations were initiated 36 h before the extreme event and simulated over a period of 144 h.

Müller et al. (2017) compared precipitation intensity from AROME-MetCoOp, ECMWF IFS, and observed



FIG. 2. The 98% percentile of the daily mean IVT $(kg\,m^{-1}\,s^{-1})$ for (a) ERA-Interim 1976–2016, (b) 30 years of EC-Earth present day, and (c) 30 years of EC-Earth future are shown.

future events had daily average precipitation of 94 and 92 mm day$^{-1}$.

A visual inspection of the IVT patterns of the selected events confirmed the initial assumption that the events were caused by atmospheric rivers, shown as the strong
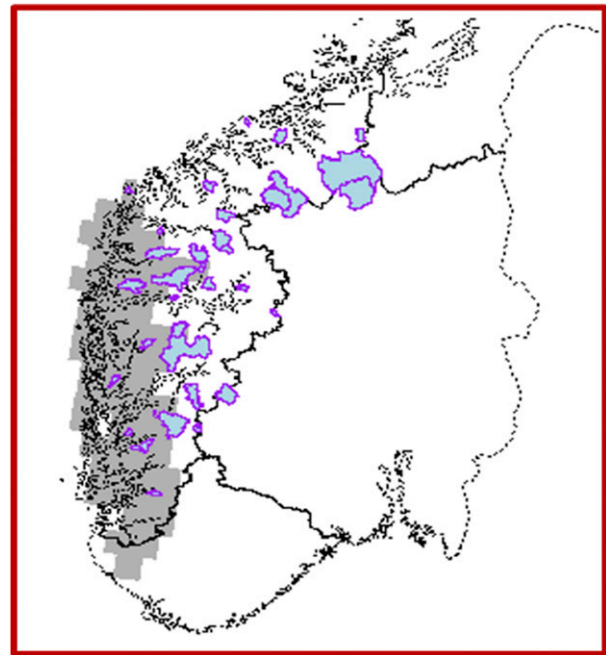
FIG. 4. The figure illustrates the EC-Earth model integrated water vapor transport (IVT) for the selected atmospheric river events: (a) pr-1, (b) pr-2, (c) fu-1, and (d) fu-2. The blue box indicates the AROME-MetCoOp domain, whereas the red box in (a) indicates the domain used to present the results of this study and includes western Norway.

station data for an atmospheric river event in October 2014 and in addition, by using a spatial verification technique over a longer time span, they showed that ECMWF IFS predicts large-scale patterns reasonably well. There is, however, a significant advancement using the high-resolution system, which could mainly be explained by better representation of orographic precipitation forcing. In Fig. 5, the 24-h accumulated precipitation during the 2005 atmospheric river event is presented, and shows that the AROME-MetCoOp model, forced with ERA-Interim, is able to reproduce the precipitation pattern compared to the gridded observations (seNorge; Lussana et al. 2018). The seNorge data are, however, too smooth, since they use a spatial interpolation scheme in an area of complex topography in order to grid the relative sparse observational network. From the above evaluation, we find that precipitation is reasonably well represented in this study, even though precipitation amounts from AROME-MetCoOp forced with ERA-Interim are low compared to the

gridded observations, and therefore should be considered as lower limits.

The AROME-MetCoOp gridded temperature and precipitation data were prepared for the hydrological model by first aggregating the gridded data to daily time resolution, and thereafter calculating the average for each catchment. All four atmospheric river events, which each consist of 10 realizations with a duration of 6 days, were used as forcing for the hydrological model.

### d. Hydrological model

The conceptual precipitation–runoff model HBV (Hydrologiska Byråns Vattenbalansavdelning) (Bergström 1976) as described in Sælthun (1996) and Beldring (2008) and implemented in the operational Norwegian flood-forecasting service, was used to estimate the streamflow. HBV has model components describing snow, soil moisture, and groundwater processes. The model is forced with daily catchment average temperature and precipitation. Within the HBV model, each catchment is

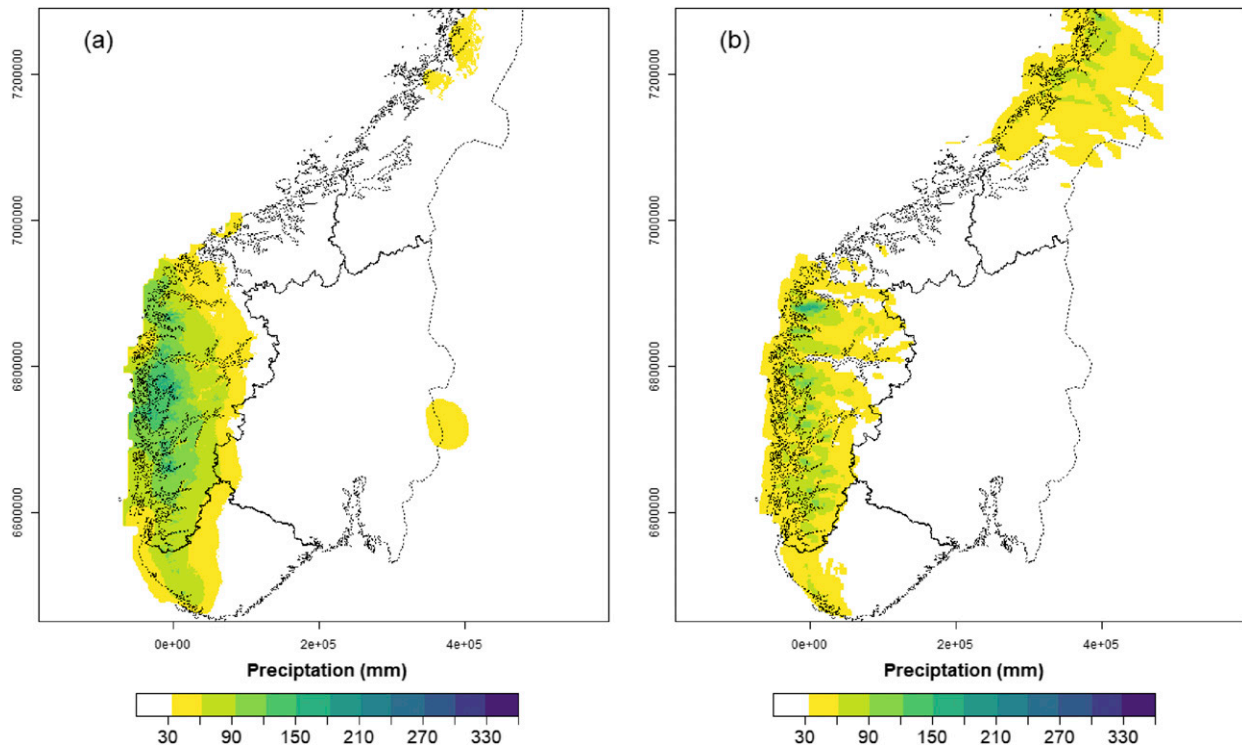FIG. 5. Precipitation accumulated over 24 h for 0600 UTC 14 Sep 2005 from (a) seNorge v2.0, interpolated observations and (b) AROME-MetCoOp, run with ERA-Interim as boundary conditions. The geo reference in seNorge $1 \times 1 \mathrm{km}^2$ grid is WGS84, UTM33 (m).

subdivided into 10 elevation zones, each elevation is chosen to yield 10 equal areas, and thereby account for vertical temperature and precipitation gradients.

All hydrological model parameters are optimized using seNorge observational precipitation and temperature data (Tveito 2002; Tveito et al. 2005; www.seNorge.no) as forcing, and streamflow observations from the hydrological database at NVE (https://www.nve.no/hydrology/) as the reference. The seNorge dataset is observational in situ data interpolated to a $1 \times 1 \mathrm{km}^2$ grid available from 1 January 1957 until today. The Nash–Sutcliffe efficiency (Nash and Sutcliffe 1970) and volume bias are used as calibration and validation metrics. The Nash–Sutcliffe efficiency (positively oriented with an optimal value of 1) averaged for all catchments is for the calibration period 0.74 with zero volume bias, and 0.71 with 2.2% volume bias for the validation period. The calibration is done for all year daily data for the period 1996–2012, whereas the validation is conducted from 1980 to 1995. This study used the same setup and parameters as the operational flood-forecasting model.

Catchments in the western region, as defined in Vormoor et al. (2016), were used to evaluate the hydrological impact of the atmospheric river events. We chose to include catchments situated outside the event selection area (as in Fig. 3) to give a better description of

the spatial impact of the events. The 37 catchments are part of the operational flood-forecasting service for Norway and most of them are located in unregulated or weakly regulated rivers. Figure 3 shows the location and area of these catchments, defined by the natural drainage area for a measuring point in the river (the gauging station). The upstream areas vary in size and elevation, and for western Norway, the catchments are steep and relatively small (from 3 to 2400 km²). Both elevation gradient and size can affect the timing and magnitude of a flood peaks. Most of the western catchments will reach the flood peak within one day of extreme precipitation. When the response time in a catchment is smaller than the model time steps, the model underestimates the flood peaks. However, current practice shows that we still get useful information on daily average flood sizes. In addition, we are evaluating the impact of atmospheric river events, lasting 24–72 h, as opposed to convective precipitation events where the subdaily time step is more critical.

In Norway, flood warnings are issued when floods exceed three predefined thresholds for daily average floods. Since there can be a discrepancy between the hydrological flood estimate and the observed values during floods, operational warning threshold are based on hydrological simulations using 60 years of seNorge

TABLE 1. A summary of precipitation and temperature for the selected historical years. The anomaly refers to the normal (reference period is 1961–90). The values represent statistics for the whole year (annual), season (autumn), and month (October) (https://www.yr.no/place/Norway/Sogn_og_Fjordane/Aurland/Fl%C3%A5m~124317/climate.html?spr=eng).

| Ic | Year | Annual | | | Autumn | | | October | | |
|----|------|--------|------|--------|--------|------|--------|---------|------|--------|
| | | $P$ | $T$ (°C) | $\Delta T$ (°C) | $P$ | $T$ (°C) | $\Delta T$ (°C) | $P$ | $T$ (°C) | $\Delta T$ (°C) |
| A | 2014 | 102.1% | 5.3 | 2.5 | 87.7% | 6.0 | 2.6 | 142% | 6.2 | 2.3 |
| B | 1997 | 110.9% | 3.4 | 0.8 | 96% | 2.9 | −0.3 | 115.8% | 1.3 | −2.3 |
| C | 1983 | 141.8% | 2.9 | 0.3 | 161% | 3.2 | 0.0 | 226.3% | 3.1 | −0.5 |
| D | 1960 | 65.5% | 2.9 | 0.3 | 37.3% | 3.3 | 0.1 | 22.7% | 1.9 | −1.6 |

observations as input. In this study, we used the operational warning levels as flood references, for example, the mean annual flood (RM), the 5-yr return level (R5), and the 50-yr return level (R50). To enable an easy comparison, the same flood thresholds were used for the present and the future events. Associated warning colors are yellow, orange and red; green is used to indicate streamflow below RM. We also extracted the maximum flood (Rmax) from these simulations and used Rmax as a reference maximum flood.

### e. Hydrological initial conditions and evaluation

In operational flood forecasting, the hydrological initial states are established by either assimilating the observed states, or by running the hydrological model with meteorological observations. The hydrological initial conditions describe the state of the water storages in catchments and are important for the catchment response to rain and snowmelt. For events of short duration, here less than 6 days, the hydrological initial conditions will affect simulated streamflow, and in this study, we wanted to investigate the impacts of this effect.

The HBV model has storages for soil moisture, groundwater, and snow. We selected a set of four different characteristic initial conditions that we combined with all atmospheric river events. The hydrological initial conditions were selected to represent characteristic combinations of snow and soil moisture storages, and they were established by running the hydrological model with selected years of seNorge observational data. We selected years that represent contrasting combinations of temperature (low, high) and precipitation (dry, wet). The selection basis was the October weather statistics for western Norway. Table 1 presents the October average temperature and precipitation for the four selected year (2014, 1992, 1983, and 1960), as well as the deviation from the normal period. We hereafter label the hydrological initial conditions as A, B, C, and D, respectively. Figure 6 illustrates the state of snow storage and soil moisture deficit for the four initial conditions on 24 October, calculated by a distributed HBV model (seNorge.no). The internal states from the

distributed model can differ, but only slightly, from the internal states of the catchment models. From Fig. 6 we see that hydrological initial condition B and D are relatively dry, whereas A and C are wet. The snow storage depends on past precipitation and temperature. There is little snow present in A and D, whereas larger areas are snow covered for both B and C. In the following the initial conditions will be described as SNOW or BARE (no snow), and the soil moisture deficit as WET or DRY. The four initial conditions can be described as A: BARE-WET, B: SNOW-DRY, C: SNOW-WET, and D: BARE-DRY.

Different sets of hydrological initial conditions allowed us to evaluate how future floods are affected not merely by the dynamical forcing, as precipitation and temperature, but also by the initial states, like soil moisture and snow, in the hydrological modeling. We used the spread in the flood realizations to evaluate the contribution caused by hydrological initial conditions and ensemble realizations. We recall that for each catchment and event, 40 different flood realizations were created by using input data from 10 realizations (the perturbations in the EC-Earth model) and four different hydrological initial conditions. We used relative mean absolute deviation (RMAD) around the mean of the flood realizations as a dimensionless measure of ensemble spread.

The $\text{RMAD}_e$ caused by the ensemble realizations (i.e., the meteorological forcing), was calculated as follows. First, the mean of the floods $\overline{q}_{ic}$ simulated by the 10 ensemble realization, was calculated for each hydrological initial condition separately and used as the central tendencies. Thereafter, the relative deviation around each $\overline{q}_{ic}$ caused by the ensemble realizations was calculated. Finally, the average of the four RMAD by the four hydrological initial conditions was calculated:

$$\text{RMAD}_e = \frac{1}{4}\sum_{ic=1}^{4}\frac{1}{10}\sum_{e=1}^{10}\left|\frac{q_{ic,e} - \overline{q}_{ic}}{\overline{q}_{ic}}\right|. \quad (1)$$

Similarly, the $\text{RMAD}_{ic}$ for the hydrological initial conditions were assessed by first calculating the deviation
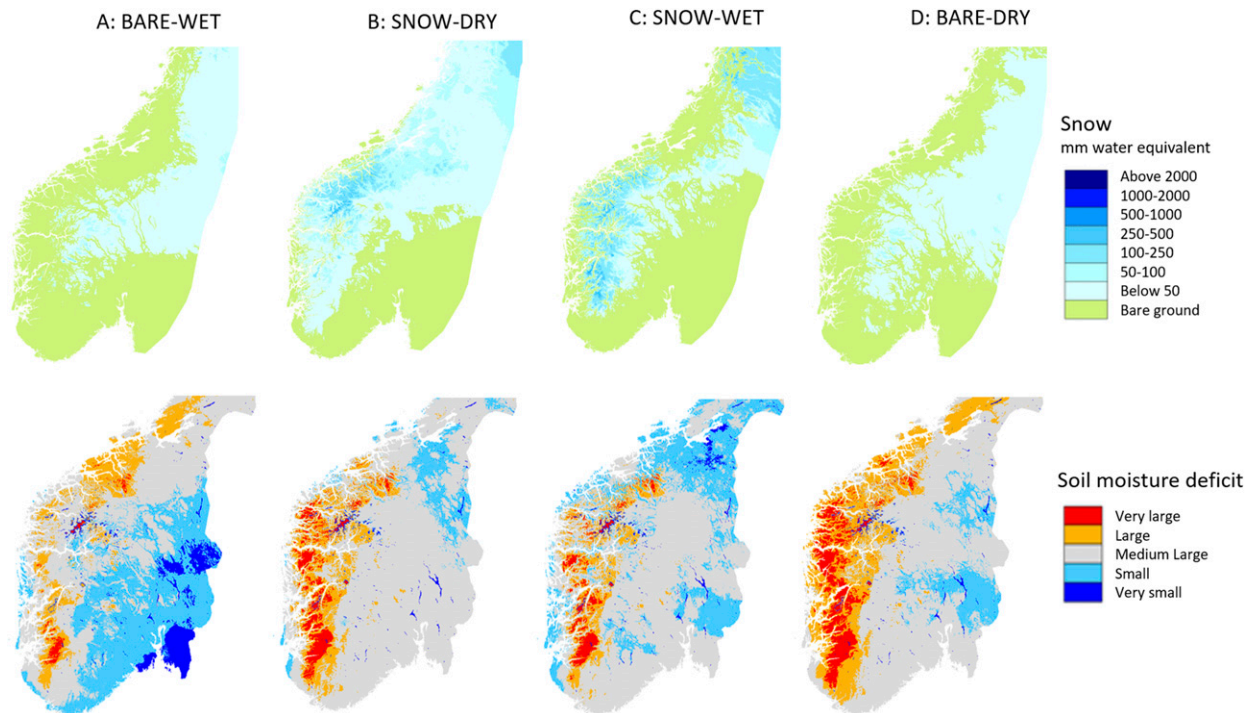
FIG. 6. (top) Snow storage and (bottom) soil moisture deficit, illustrating the four different hydrological initial conditions used in the HBV model. For snow (mm), the blue colors indicate snow, and green bare ground. For soil moisture deficit, the red color indicates very dry conditions, whereas blue indicates wet. The soil moisture deficit indicates to which percentile, defined by a reference period (1980–2010), the daily value categorizes. The 5th, 25th, 50th, 75th, and 95th percentiles of the reference period define the categories from "very small/wet" to "very large/dry."

around each ensemble realizations due to the hydrological initial conditions, using the mean of the ensemble realizations $\overline{q}_e$ as the central tendency:

$$\text{RMAD}_{\text{ic}} = \frac{1}{10}\sum_{e=1}^{10}\frac{1}{4}\sum_{\text{ic}=1}^{4}\left|\frac{q_{\text{ic},e} - \overline{q}_e}{\overline{q}_e}\right|. \quad (2)$$

We calculated the RMAD only at the time of the flood peak for each event. By comparing $\text{RMAD}_e$ to $\text{RMAD}_{\text{ic}}$ in a scatterplot, we can assess the importance of the ensemble realizations and the hydrological initial conditions to the spread in simulated flood sizes. Both $\text{RMAD}_e$ and $\text{RMAD}_{\text{ic}}$ will have a value between 0 and 1.

## 3. Results and discussion

### a. The precipitation events

Figure 7 illustrates the spatial distribution of 132 h accumulated precipitation during the four atmospheric river events simulated with AROME-MetCoOp and averaged over the 10 realizations. The future (fu) events fu-1 (Fig. 7c) and fu-2 (Fig. 7d) have more precipitation, and especially coastal precipitation is higher compared to present (pr) events pr-1 (Fig. 7a) and pr-2 (Fig. 7b).

The IVT of both future events are higher (Figs. 4c,d), and are located more to the north, compared to the present events. The higher moisture content in the future events means that precipitation will fall out at lower elevation. (i.e., closer to the coast). This finding is supported by Sandvik et al. (2018) conducting a study for the Norwegian west coast. They find that an increase in temperature (as in a future climate) causes a larger increase in near-coastal precipitation compared to more inland precipitation.

Pr-1 has overall less precipitation than the other events. Whereas pr-2, which has below zero temperatures for larger parts of the domain, has some precipitation peaks, located at the highest elevations, and over glaciered areas. The peaks in pr-2 seem to be induced by the downscaling, but do not affect flood magnitudes, since the temperatures are well below zero.

### b. Flood evaluation

#### 1) FLOOD WARNING LEVELS

For all catchments and realizations, we determined the floods exceedance of operational flood warning thresholds. Figure 8 gives an overview of catchments
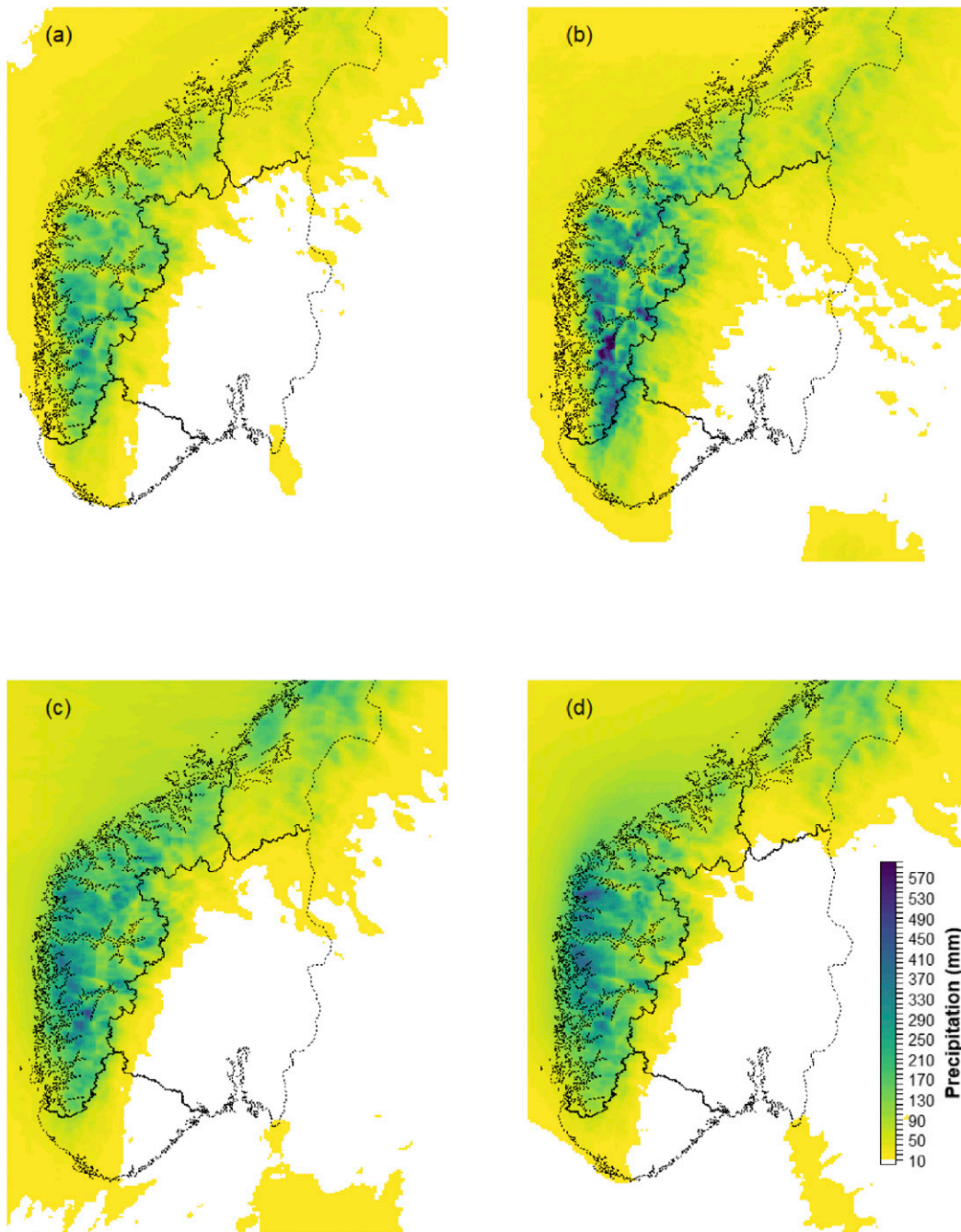
FIG. 7. Accumulated precipitation (132 h) for present events (a) pr-1 and (b) pr-2 and future events (c) fu-1 and (d) fu-2. All values are the average of 10 realizations of AROME-MetCoOp precipitation.

with no floods (green) and catchments where at least one flood realization reached the warning levels of mean annual flood (yellow), 5-yr flood (orange), and 50-yr flood (red), which are the official warning levels in Norway. In each colored cell, the number indicates the percentage of realizations (10 members in each cell) that exceeded the red level. The columns in Fig. 8 represent both present and future events combined with the four different initial conditions, A–D (Table 1). Each row shows the results for one catchment. The last row, however, represents the percentage of R50 exceedances that includes all catchments and is therefore a summary of the combined effect of initial conditions and events on floods over the larger area. Figure 8 highlights that the

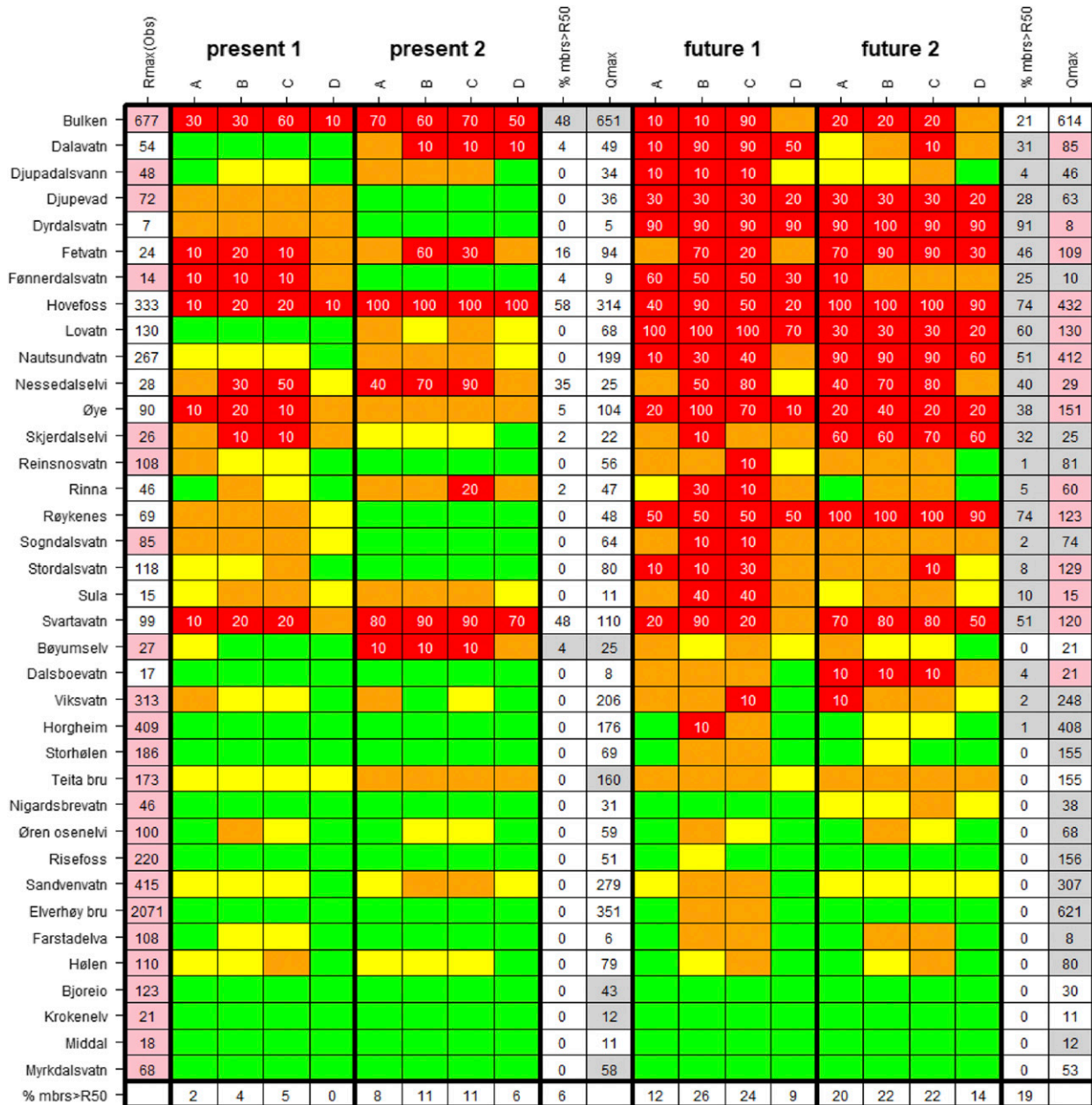| | Rmax (Obs) | present 1 A | B | C | D | present 2 A | B | C | D | % mbrs>R50 | Qmax | future 1 A | B | C | D | future 2 A | B | C | D | % mbrs>R50 | Qmax |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bulken | 677 | 30 | 30 | 60 | 10 | 70 | 60 | 70 | 50 | 48 | 651 | 10 | 10 | 90 | | 20 | 20 | 20 | | 21 | 614 |
| Dalavatn | 54 | | | | | | 10 | 10 | 10 | 4 | 49 | 10 | 90 | 90 | 50 | | | 10 | | 31 | 85 |
| Djupadalsvann | 48 | | | | | | | | | 0 | 34 | 10 | 10 | 10 | | | | | | 4 | 46 |
| Djupevad | 72 | | | | | | | | | 0 | 36 | 30 | 30 | 30 | 20 | 30 | 30 | 30 | 20 | 28 | 63 |
| Dyrdalsvatn | 7 | | | | | | | | | 0 | 5 | 90 | 90 | 90 | 90 | 90 | 100 | 90 | 90 | 91 | 8 |
| Fetvatn | 24 | 10 | 20 | 10 | | | 60 | 30 | | 16 | 94 | | 70 | 20 | | 70 | 90 | 90 | 30 | 46 | 109 |
| Fønnerdalsvatn | 14 | 10 | 10 | 10 | | | | | | 4 | 9 | 60 | 50 | 50 | 30 | 10 | | | | 25 | 10 |
| Hovefoss | 333 | 10 | 20 | 20 | 10 | 100 | 100 | 100 | 100 | 58 | 314 | 40 | 90 | 50 | 20 | 100 | 100 | 100 | 90 | 74 | 432 |
| Lovatn | 130 | | | | | | | | | 0 | 68 | 100 | 100 | 100 | 70 | 30 | 30 | 30 | 20 | 60 | 130 |
| Nautsundvatn | 267 | | | | | | | | | 0 | 199 | 10 | 30 | 40 | | 90 | 90 | 90 | 60 | 51 | 412 |
| Nessedalselvi | 28 | | 30 | 50 | | 40 | 70 | 90 | | 35 | 25 | | 50 | 80 | | 40 | 70 | 80 | | 40 | 29 |
| Øye | 90 | 10 | 20 | 10 | | | | | | 5 | 104 | 20 | 100 | 70 | 10 | 20 | 40 | 20 | 20 | 38 | 151 |
| Skjerdalselvi | 26 | | 10 | 10 | | | | | | 2 | 22 | | 10 | | | 60 | 60 | 70 | 60 | 32 | 25 |
| Reinsnosvatn | 108 | | | | | | | | | 0 | 56 | | | 10 | | | | | | 1 | 81 |
| Rinna | 46 | | | | | | | 20 | | 2 | 47 | | 30 | 10 | | | | | | 5 | 60 |
| Røykenes | 69 | | | | | | | | | 0 | 48 | 50 | 50 | 50 | 50 | 100 | 100 | 100 | 90 | 74 | 123 |
| Sogndalsvatn | 85 | | | | | | | | | 0 | 64 | | 10 | 10 | | | | | | 2 | 74 |
| Stordalsvatn | 118 | | | | | | | | | 0 | 80 | 10 | 10 | 30 | | | | 10 | | 8 | 129 |
| Sula | 15 | | | | | | | | | 0 | 11 | | 40 | 40 | | | | | | 10 | 15 |
| Svartavatn | 99 | 10 | 20 | 20 | | 80 | 90 | 90 | 70 | 48 | 110 | 20 | 90 | 20 | | 70 | 80 | 80 | 50 | 51 | 120 |
| Bøyumselv | 27 | | | | | 10 | 10 | 10 | | 4 | 25 | | | | | | | | | 0 | 21 |
| Dalsboevatn | 17 | | | | | | | | | 0 | 8 | | | | | 10 | 10 | 10 | | 4 | 21 |
| Viksvatn | 313 | | | | | | | | | 0 | 206 | | | 10 | | 10 | | | | 2 | 248 |
| Horgheim | 409 | | | | | | | | | 0 | 176 | | 10 | | | | | | | 1 | 408 |
| Storhølen | 186 | | | | | | | | | 0 | 69 | | | | | | | | | 0 | 155 |
| Teita bru | 173 | | | | | | | | | 0 | 160 | | | | | | | | | 0 | 155 |
| Nigardsbrevatn | 46 | | | | | | | | | 0 | 31 | | | | | | | | | 0 | 38 |
| Øren osenelvi | 100 | | | | | | | | | 0 | 59 | | | | | | | | | 0 | 68 |
| Risefoss | 220 | | | | | | | | | 0 | 51 | | | | | | | | | 0 | 156 |
| Sandvenvatn | 415 | | | | | | | | | 0 | 279 | | | | | | | | | 0 | 307 |
| Elverhøy bru | 2071 | | | | | | | | | 0 | 351 | | | | | | | | | 0 | 621 |
| Farstadelva | 108 | | | | | | | | | 0 | 6 | | | | | | | | | 0 | 8 |
| Hølen | 110 | | | | | | | | | 0 | 79 | | | | | | | | | 0 | 80 |
| Bjoreio | 123 | | | | | | | | | 0 | 43 | | | | | | | | | 0 | 30 |
| Krokenelv | 21 | | | | | | | | | 0 | 12 | | | | | | | | | 0 | 11 |
| Middal | 18 | | | | | | | | | 0 | 11 | | | | | | | | | 0 | 12 |
| Myrkdalsvatn | 68 | | | | | | | | | 0 | 58 | | | | | | | | | 0 | 53 |
| % mbrs>R50 | | 2 | 4 | 5 | 0 | 8 | 11 | 11 | 6 | 6 | | 12 | 26 | 24 | 9 | 20 | 22 | 22 | 14 | 19 | |

FIG. 8. An overview of warning levels for the combination of events and initial conditions (columns A–D) for all 37 catchments (rows). The colors indicate the flood warning level exceeded, while the numbers within the red cells indicate the percentage of ensemble members reaching a red warning level for that specific combination of event and initial condition. The total red warning levels reached for each catchment (% mbrs > R50), and the maximum flood level ($m^3\,s^{-1}$; Qmax) estimated for any of the events is presented in columns. The gray colored cells indicate the highest Qmax value for each catchment comparing future and present events, whereas pink colored cells indicate the highest streamflow value comparing Qmax to Rmax ($m^3\,s^{-1}$). The last row represents the percentage of R50 exceedances for all catchments and is therefore a summary of the combined effect of initial conditions and events on floods for the larger area.

red warning level is reached for more catchments in the future compared to the present events, from 23 to 11, respectively. Moreover, for the future events, a higher number of the realizations reached the highest warning level (% mbrs > R50), totally 18.9% in the future and 6.1% in the present climate (see % mbrs > R50 in columns 10 and 20 in Fig. 8). By dividing the total number of realizations from Fig. 8 by the number of catchments impacted summarized from Table 2, we find that on average there are 3.8 realizations per red warning in the

TABLE 2. Total number of catchments where streamflow exceeds red warning levels, indexed by initial conditions and events. Initial conditions (ic) are based the soil moisture and snow conditions from Fig. 7, which refers to Oct data for selected years (Table 1).

| Ic event | A: BARE-WET | B: SNOW-DRY | C: SNOW-WET | D: BARE-DRY | Red/event |
|---|---|---|---|---|---|
| Pr-1 | 5 | 7 | 7 | 2 | 21 |
| Pr-2 | 6 | 7 | 8 | 5 | 26 |
| Fu-1 | 13 | 19 | 20 | 7 | 59 |
| Fu-2 | 14 | 13 | 15 | 9 | 51 |
| Red/ic | 38 | 46 | 50 | 23 | 157 |

present climate, and 5.1 realizations per red warning in the future climate. Overall, the results show that more catchments and more realizations reached a red warning level in the future events.

The flood magnitude is important for the assessment of flood impact, since the magnitude directly relates to damage potential. To evaluate the possible maximum impact from the present and the future realizations, we select the highest flood value ($Q_{max}$; $m^3 s^{-1}$) for each catchment; hence, each $Q_{max}$ was selected from two events and 40 realizations per event. $Q_{max}$ representing the present and the future climate (Fig. 8) were compared with each other and to the reference maximum floods (Rmax). Note that Rmax is the highest daily flood value estimated running the hydrological model with 60 years of seNorge temperature and precipitation observational data and is not constrained to atmospheric river events. The evaluation shows that $Q_{max}$ for the future events is higher than $Q_{max}$ for the present events for 22 of 23 catchments that reached a red warning for the future events (gray or pink colored cells in the last column in Fig. 8). The future highest floods are larger than the reference floods (Rmax) for 14 of the same 23 catchments, whereas none of the present climate floods exceeds Rmax.

In summary, these results show that for the future atmospheric river events, more catchments will reach a red warning level during the same event. For these catchments, the daily flood peaks will be higher, and more realizations will reach a red warning level, indicating that extreme floods coinciding in multiple catchments are more likely in the future. It is important to note that the results are conservative and might underestimate the flood risk for two reasons: (i) we are using a moderate RCP4.5 scenario and (ii) the extreme precipitation of the events seems to be low compared to the gridded interpolated observations (Fig. 5).

### 2) HYDROLOGICAL INITIAL CONDITIONS

Figure 8 moreover holds information on the sensitivity of floods to initial conditions, expressed by the change in warning colors with the initial conditions A–D for several of the catchments. Table 2 summarizes the catchments from Fig. 8 exceeding a red level. Snow in

the catchments clearly contribute to increased floods, the total number of catchments exceeding red level for all events was 96 for SNOW (B and C) versus 61 for BARE (A and D). The difference between catchments exceeding the red level for SNOW-DRY (46) and SNOW-WET (50) is relatively small; illustrating that snow is the most important hydrological initial condition in this area. The soil moisture is more important when there is little or no snow available, for BARE-WET 38 catchments exceed the red warning level compared to BARE-DRY where 23 exceeded, which indicate an increase in number of exceedances of 65% when conditions are wet. By contrast, when snowpack is significant, there is an increase of only 11% under wet versus dry soil moisture conditions.

The initial conditions for a future scenario are more likely to be similar to A (BARE-WET), since climate projections indicate a warmer and wetter west coast climate (Baatsen et al. 2015; Hanssen-Bauer et al. 2017). However, variabilities in weather in the future should also be anticipated, and the exceptional October 2018 flood event in western Norway exemplified (see section 1) the importance of the variability in temperature and precipitation during autumn. The presence of snow that is available for melt followed by a warm and moist atmospheric river event (extreme precipitation) has historically shown to cause some of the most severe flood impacts (Roald 2008). Figure 8, thus, illustrates that accounting for different hydrological initial conditions in the climate–hydrological modeling chain adds value to the flood impact evaluation. The antecedent conditions and the memory of the catchment, for example, groundwater, soil moisture, water levels in lakes, and snow, are all important for the catchment response to extreme precipitation.

### 3) SPATIAL FLOOD IMPACT

Historical atmospheric river events clearly show that the spatial impact and severity of the events varies. Which and to what extent catchments are affected by an atmospheric river depends on several factors: the atmospheric conditions (e.g., the placements of pressure systems), the extent and duration of the atmospheric river, the moisture content, the temperature, and the intensity and duration of precipitation during the event.
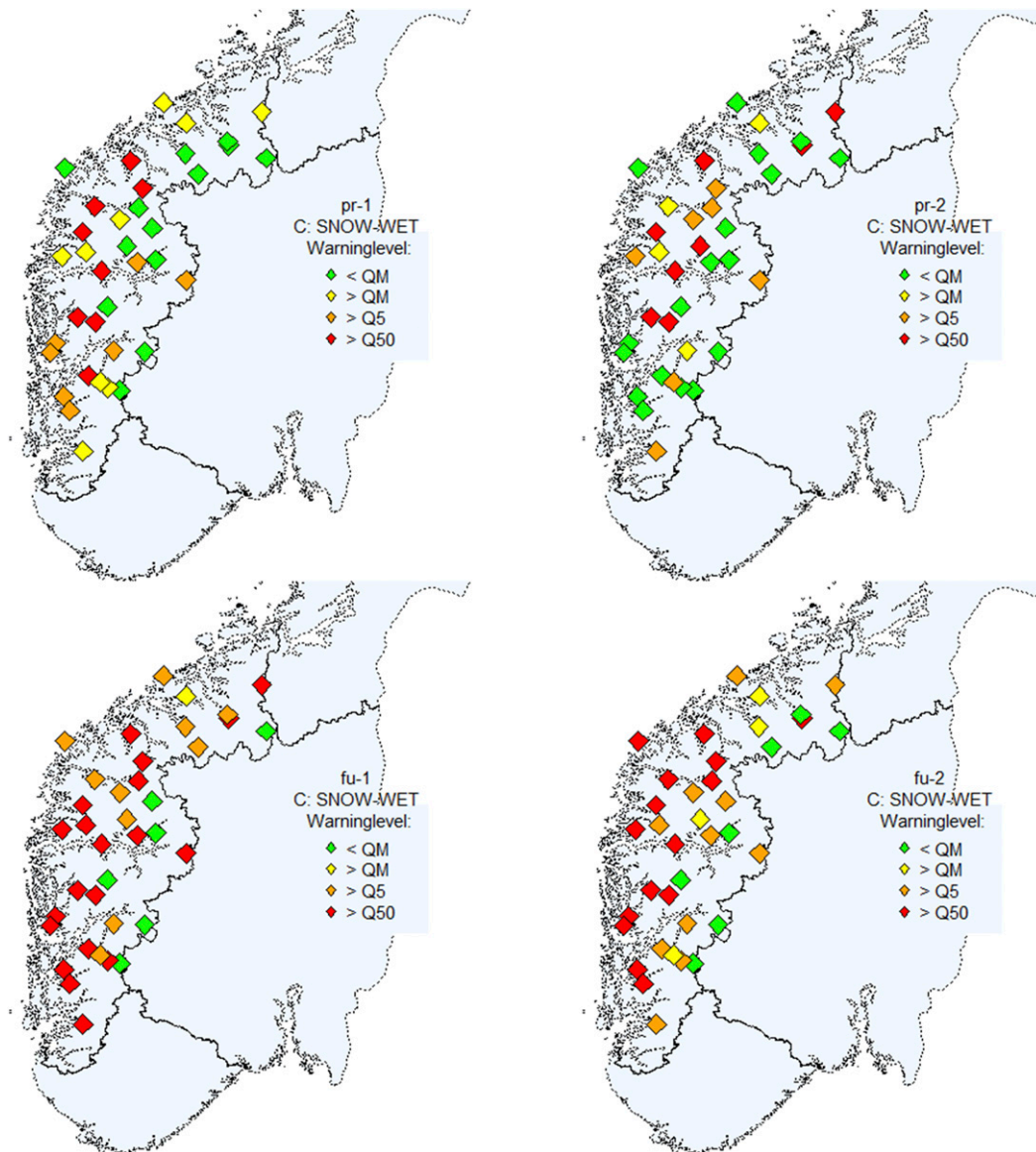
FIG. 9. The spatial distribution of the warning levels for catchments in western Norway, SNOW-WET (C) used to establish initial conditions. The highest warning level reached for each catchment is presented for all events: (top) pr-1 and pr-2 (present climate) and (bottom) fu-1 and fu-2 (future climate). Yellow (RM), orange (R5), and red (R50) diamonds define floods exceeding the warning thresholds, whereas green diamonds indicate no floods.

We used the maximum flood level reached at each catchment to visualize the spatial impact of the four atmospheric river events, presented by separate maps in Fig. 9. SNOW-WET (C) was used as hydrological initial condition for all events, since it was the hydrological initial condition that resulted in the highest number of catchments exceeding the red flood level in both present and future climate (Table 2).

Figure 9 shows that the future events affect more catchments closer to the coast than the present events. This seems to be in line with Sandvik et al. (2018). They

used an idealized temperature perturbation to study the effect of temperature increase on precipitation and found that near-coastal, high-elevation areas experienced the highest increase in extreme precipitation. Apart from higher future floods for the coastal catchments, there are no other clear patterns to be seen in Fig. 9. Not all catchments were affected by all events, mainly because each landfall is unique, and some areas are missed by the most intense precipitation. In a few catchments, however, a temperature below freezing at higher elevations cause precipitation to classify as snow

in parts of catchment, and for these catchments high precipitation does not generate floods. Event pr-2 is relatively cold, which resulted in fewer floods even if the precipitation was high (Fig. 4b).

We find that Fig. 9 demonstrates an advantage of the event-based approach by providing a plausible image of the possible impact of single events, using a well-known modeling chain and flood exceedance levels. This way, it is easier to present how future atmospheric river events, that contain more precipitation due to a warmer climate, have a higher impact over a larger spatial area with severe floods occurring in multiple catchments within a short time span.

### c. Sensitivity to ensemble spread and hydrological initial conditions

The spatial differences in precipitation combined with the hydrological initial conditions will determine how floods evolve and unfold in different catchments. To visualize the different flood impacts and the range of possible outcomes, detailed results for six catchments are presented in Fig. 10, with catchment characteristics shown in Table 3. Each boxplot in Fig. 10 represents 40 realizations and reveals the range in streamflow estimates due to the combination of meteorological ensemble realizations and the hydrological initial conditions.

Øye and Lovatn are located north in the study domain and outside the event selection area. Despite a spatial proximity (35-km distance, but in different fjords), they reveal different streamflow response to the same atmospheric river events. The spread in the streamflow response to meteorological forcing and hydrological initial conditions are larger for Øye than Lovatn. Nautsundvatn, Hovefoss, and Røykenes are all near-coastal catchments with similarities in catchment characteristics. The high future floods are representative for the increased near-coastal precipitation seen in fu-1 and fu-2 (Fig. 4). The three catchments all show a similar response to the precipitation, with respect to both magnitudes and timing. Røykenes and Bulken, on the other hand, illustrate contrasting flood responses. Røykenes is an example of a small and steep catchment with a quick streamflow response to precipitation, whereas for Bulken, with a larger catchment area, streamflow rises slower and the high streamflow lasts longer. Bulken is positioned central within the event selection area and is affected by high floods in both the present and future simulations.

The large spread in the flood estimates (boxplots in Fig. 10) indicates sensitivity to the hydrological initial conditions and/or the temperature and precipitation forcing. The large spread demonstrates that an ensemble

of realizations is important to capture the highest floods that pose the highest damage potential. The spread can be attributed both to the ensemble realization and to the hydrological initial conditions. In Fig. 11, using RMAD as a measure of spread [see Eqs. (1) and (2)], we find that for most events and catchments, the ensemble realizations contribute most to the spread, however, for some events and catchments, the hydrological initial conditions contribute more. Figure 11 does not reveal any clear pattern that would explain whether spread caused by ensembles realizations or hydrological initial conditions were more important for any catchments, however, for all catchments in pr-1 and fu-2 the ensemble spread has the highest contribution to the spread in the flood estimates. For the catchments and events in Fig. 11, we find that ensemble realizations can change the streamflow estimates by over 60% (Øye, pr-1), whereas the initial conditions can alter the estimates by nearly 30% (Nautsundvatn, pr-1). Mostly the contribution to spread caused by both hydrological initial conditions and ensemble realizations are between 10% and 20%. An evaluation that included all catchments gave similar results.

### d. Final remarks on the event-based approach

In this study, we used a small sample size, which in a traditional setup would mean that we could not draw any conclusions to whether our results are due to natural variability or from climate change. We present 40 EC-Earth realizations per catchment, whereof 20 are set in a future and 20 in a present climate. All realizations are from one climate model, compared to multimodel ensembles used in most climate projection studies (e.g., Sillmann et al. 2013), and our results represent only one emission scenario (RCP4.5). Nevertheless, we believe that the event-based approach provides supplementary information about future flood risk in western Norway. First, our results are in accordance with previous studies of climate change in this region. They show increased precipitation in western Norway independent of model or emission scenario (e.g., Hanssen-Bauer et al. 2017), and our events are hence representative for these scenarios. Moreover, Whan et al. (2020) show that the extreme precipitation and atmospheric river statistics in the present climate EC-Earth simulations is similar to the observed relationship (ERA-Interim), and that in a future climate the frequency of atmospheric rivers and the intensity of each event increases. We further know that the high-resolution EC-Earth model is suitable to model the precipitation onto the topography of the west coast of Norway, and we can assume that the events are plausible under both the present and the future climate. In addition, we used the operational AROME-MetCoOp
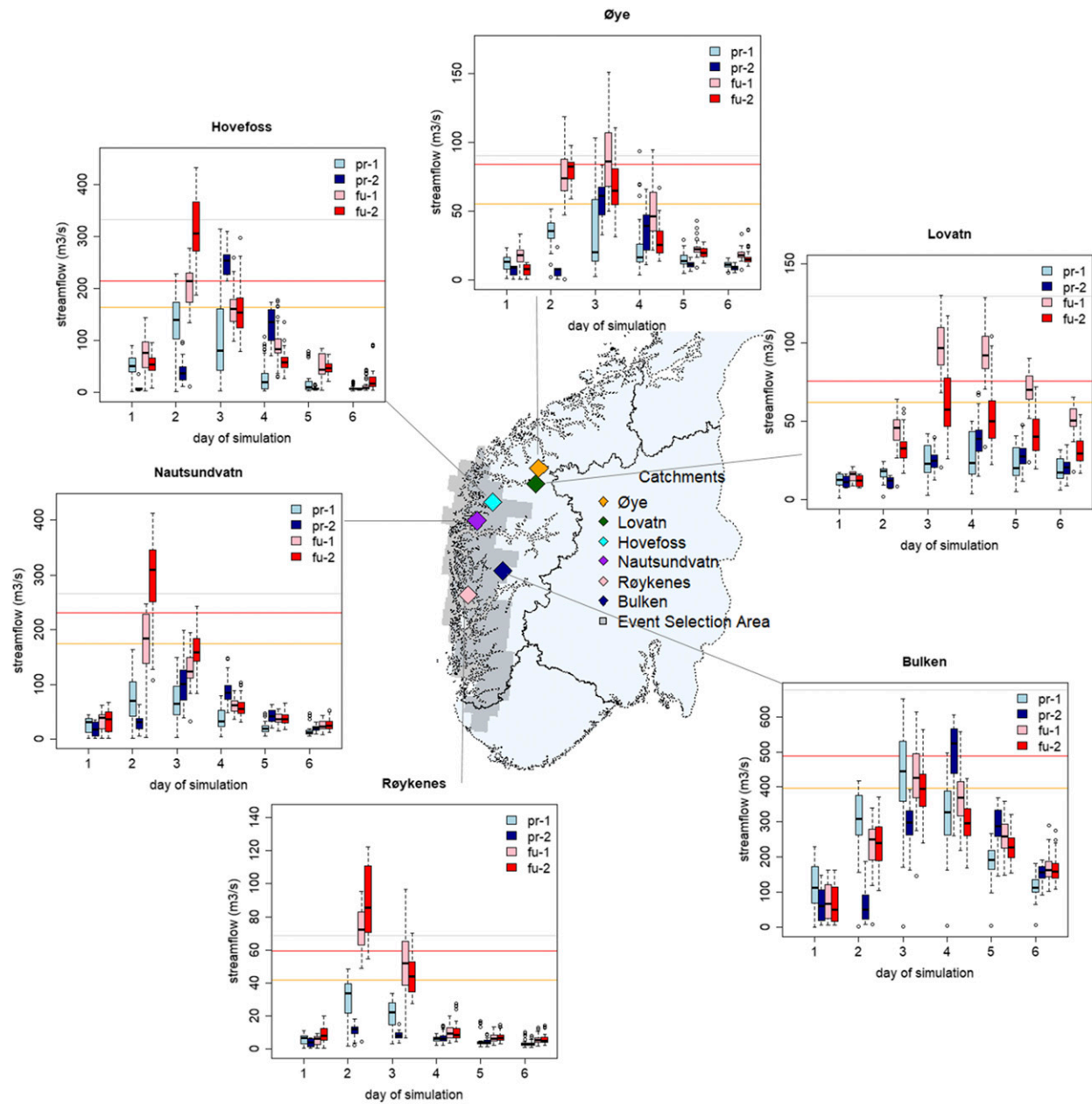
FIG. 10. Hydrographs of boxplots showing the 4 × 10 estimated floods for six selected catchments. Present events are colored blue: pr-1 is light blue, pr-2 is blue, and future events are colored green: fu-1 is light-green (pink), and fu-2 is green (red). The horizontal orange and red lines are the thresholds for 5- and 50-yr floods, while the gray line is the reference maximum floods (Rmax), which is calculated from 60 years of seNorge data for respective catchments; the *y* axis is streamflow and *x* axis is days of the simulated event. Each box contains 40 values, the box itself defines the 25th–75th percentile, and the whiskers max 1.5 times the box extension with any values outsides indicated as circular outliers.

weather prediction model, which improves the representation of extreme precipitation events (Müller et al. 2017), and the operational HBV models to finally estimate the floods. We kept the physical consistency of the variables throughout the modeling chain and could hence provide physically plausible flood events. The similarities to the operational modeling chain, makes the results easily accessible to both flood management and scientist.

## 4. Summary and conclusions

We compared possible flood events in a future warmer climate to those in a present climate by analyzing four

TABLE 3. Catchment characteristics for catchments in the case study. Area is the catchment area upstream the gauging station, annual precip is the annual catchment average precipitation, lake fraction is the lake covered area in percent of total area, steepness is (max elev − min elev)/area, min elev is minimum catchment elevation, and max elev is maximum catchment elevation.

| Catchment | Area (km²) | Annual precip (mm) | Lake fraction (%) | Steepness (m km⁻²) | Min elev (m MSL) | Max elev (m MSL) |
|---|---|---|---|---|---|---|
| Øye | 138 | 976 | 0.26 | 12.25 | 147 | 1848 |
| Lovatn | 234 | 2041 | 4.5 | 8.59 | 52 | 2071 |
| Hovefoss | 234 | 2796 | 0.45 | 6.18 | 20 | 1469 |
| Nautsundvatn | 219 | 3043 | 2.16 | 3.92 | 45 | 904 |
| Bulken | 1092 | 2037 | 0.87 | 1.42 | 47 | 1602 |
| Røykenes | 50 | 3176 | 2.24 | 18.12 | 53 | 960 |

representative extreme precipitation events associated with atmospheric rivers reaching the west coast of Norway. In addition, we introduced four different hydrological initial conditions, established using historical data, to evaluate the importance of antecedent weather on floods, and thereby uncertainties introduced by the hydrological initial conditions. We followed an event-based approach that particularly facilitates the use of a modeling chain similar to the operational weather- and flood-forecasting chain. This approach enabled us to easier interpret the possible impact of such events and makes the results accessible to both a scientific and a management audience.

This study shows the following:

- The future atmospheric river events affect larger areas compared to the present events. We found that the future events caused floods in more catchments, and the floods were at a higher magnitude.
- As each event is unique, the landfall of the atmospheric rivers is difficult to determine beforehand. Hence, applying hydrological modeling to several catchments within a larger area was important to capture the total spatial flood impact of each event.
- Hydrological initial conditions, for example, soil saturation and snow storage, affected the flood magnitude for most of the catchments and for all events. Available snow increased the number of catchments in flood due to the added contribution from snowmelt, whereas dry and snow free conditions reduced the number of catchments in flood. Snow storage was the most important initial condition, showing that an atmospheric river event following a period with snow accumulation has the highest damage potential.
- The ensemble realizations, representing different precipitation intensities, were the major contributor to spread in the flood estimates for most catchments, however, for one quarter of catchments and events, the hydrological initial conditions exhibited an important contribution to the spread.

We acknowledge that this event-based approach, which contains results from a few plausible events, based on only

one climate model and only one emission scenario, will not provide enough data to provide any probabilities for future floods. We do, however, provide high-resolution modeling, well adopted to describe atmospheric rivers in both the present and the future climate. Further event-based studies could include events simulated in different emission scenarios like the RCP8.5, which projects even more extreme precipitation for western Norway.

The benefit of the event-based approach is that the events are easier to visualize and to communicate, than for example multimodel ensemble probabilities. The event-based approach is therefore a useful supplementary to raise the awareness of possible future impact, caused by physically plausible extreme events. The chosen approach
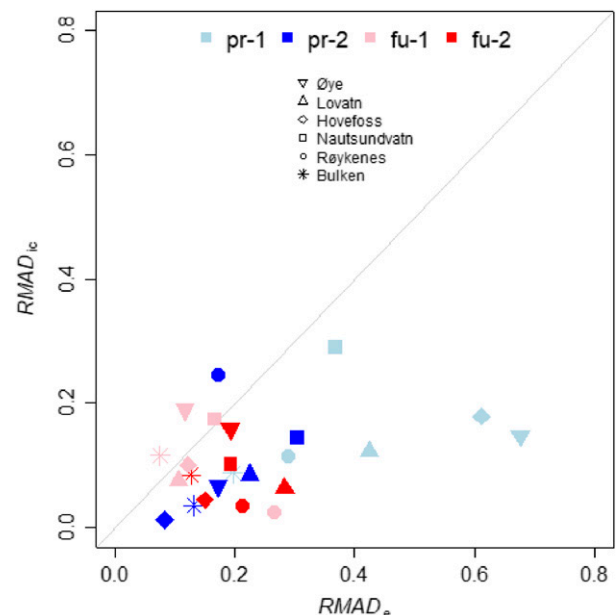


FIG. 11. The same catchments as in Fig. 10, but here the contributions to spread seen in the boxplots are separated to that caused by the ensemble realizations representing the meteorological forcing (e) and the hydrological initial conditions (ic). Relative mean absolute deviation caused by the ensemble realizations ($RMAD_e$) is on the x axis, and the relative mean absolute deviation caused by the hydrological initial conditions ($RMAD_{ic}$) is on the y axis.

operates very closely to current operational procedures, using a strategic modeling chain ranging from GCMs over RCMs to hydrological models; this makes it accessible to both the scientific community and a management audience. Floods, which are often accompanied by landslides, are demanding situations for rescue and emergency preparedness and cause high economic and social losses in affected areas. By providing information that implies that future events most likely will affect more catchments, and hence involve more municipalities simultaneously, such information can reveal some future challenges for, for example, municipalities and railway and road authorities. These future challenges underline the need for joint preparedness across, for example, community boarders for exposed and vulnerable areas.

Data availability statement: Data are available upon reasonable request from NVE (HBV), MetNorway (AROME-MetCoOp), and KNMI (EC-Earth). SeNorge data are available from the free meteorological data portal (https://www.met.no/en/free-meteorological-data/Download-services), and discharge measurements are available from the hydrological database at NVE (https://www.nve.no/hydrology/).

## REFERENCES

Azad, R., and A. Sorteberg, 2017: Extreme daily precipitation in coastal western Norway and the link to atmospheric rivers. *J. Geophys. Res. Atmos.*, **122**, 2080–2095, https://doi.org/10.1002/2016JD025615.

Baatsen, M., R. J. Haarsma, A. J. Van Delden, and H. de Vries, 2015: Sever Autumn storms in future Western Europe with a warmer Atlantic Ocean. *Climate Dyn.*, **45**, 949–964, https://doi.org/10.1007/s00382-014-2329-8.

Beldring, S., 2008: Distributed element water balance model system. NVE Rep. 4, 42 pp., http://publikasjoner.nve.no/report/2008/report2008_04.pdf.

Benedict, I., K. Ødemark, T. Nipen, and R. Moore, 2019: Large-scale flow patterns associated with extreme precipitation and atmospheric rivers over Norway. *Mon. Wea. Rev.*, **147**, 1415–1428, https://doi.org/10.1175/MWR-D-18-0362.1.

Berghuijs, W. R., S. Harrigan, P. Molnar, L. J. Slater, and J. W. Kirchner, 2019: The relative importance of different flood-generating mechanisms across Europe. *Water Resour. Res.*, **55**, 4582–4593, https://doi.org/10.1029/2019WR024841.

Bergström, S., 1976: Development and application of a conceptual runoff model for Scandinavian catchments. SMHI RHO Rep. 7, 134 pp.

Bintanja, R., C. Severijns, R. Haarsma, and W. Hazeleger, 2014: The future of Antarctica's surface winds simulated by a high-resolution global climate model: 1. Model description and validation. *J. Geophys. Res. Atmos.*, **119**, 7136–7159, https://doi.org/10.1002/2013JD020847.

Caroletti, G. N., and I. Barstad, 2010: An assessment of future extreme precipitation in western Norway using a linear model. *Hydrol. Earth Syst. Sci.*, **14**, 2329–2341, https://doi.org/10.5194/hess-14-2329-2010.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Dettinger, M., 2011: Climate change, atmospheric rivers, and floods in California – A multimodel analysis of storm frequency and magnitude changes. *J. Amer. Water Resour. Assoc.*, **47**, 514–523, https://doi.org/10.1111/j.1752-1688.2011.00546.x.

Espinoza, V., D. E. Waliser, B. Guan, D. A. Lavers, and F. M. Ralph, 2018: Global analysis of climate change projection effects on atmospheric rivers. *Geophys. Res. Lett.*, **45**, 4299–4308, https://doi.org/10.1029/2017GL076968.

Haarsma, R. J., W. Hazeleger, C. Severijns, H. Vries, A. Sterl, R. Bintanja, G. J. Oldenborgh, and H. W. Brink, 2013: More hurricanes to hit Western Europe due to global warming. *Geophys. Res. Lett.*, **40**, 1783–1788, https://doi.org/10.1002/grl.50360.

Hanssen-Bauer, I., and Coauthors, 2017: Climate in Norway 2100 - A knowledge base for climate adaption. NCCS Rep. 1/2017, 48 pp.

Hazeleger, W., and Coauthors, 2010: EC-Earth: A seamless earth-system prediction approach in action. *Bull. Amer. Meteor. Soc.*, **91**, 1357–1364, https://doi.org/10.1175/2010BAMS2877.1.

——, B. J. J. M. van den Hurk, E. Min, G. J. van Oldenborgh, A. C. Petersen, D. A. Stainforth, E. Vasileiadou, and L. A. Smith, 2015: Tales of future weather. *Nat. Climate Change*, **5**, 107–113, https://doi.org/10.1038/nclimate2450.

Howarth, C., D. Viner, S. Dessai, C. Rapley, and A. Jones, 2017: Enhancing the contribution and role of practitioner knowledge in the Intergovernmental Panel on Climate Change (IPCC) Working Group (WG) II process: Insights from UK workshops. *Climate Serv.*, **5**, 3–10, https://doi.org/10.1016/j.cliser.2017.04.003.

Iden, K., K. Isaksen, S. Kristiansen, and H. Szewczyk-Bartnicka, 2005: Weather in Norway—Climatological monthly overview for September 2005 (in Norwegian). Meteorological Institute, Oslo, Norway, 18 pp., https://www.met.no/publikasjoner/met-info/met-info-2005.

IPCC, 2010: IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections: Meeting Report. T. Stocker et al., Eds., IPCC, 127 pp., https://www.ipcc.ch/site/assets/uploads/2018/05/expert-meeting-assessing-multi-model-projections-2010-01.pdf.

Kobierska, F., K. Engeland, and T. Thorarinsdottir, 2018: Evaluation of design flood estimates – A case study for Norway. *Hydrol. Res.*, **49**, 450–465, https://doi.org/10.2166/nh.2017.068.

Langsholt, E., L. A. Roald, E. Holmqvist, and A. Fleig, 2015: Flommen på Vestlandet oktober 2014 (in Norwegian). Norwegian Water Resources and Energy Directorate, Rep. 112015, 69 pp., http://publikasjoner.nve.no/rapport/2015/rapport2015_11.pdf.

Lavers, D. A., and G. Villarini, 2013: The nexus between atmospheric rivers and extreme precipitation across Europe. *Geophys. Res. Lett.*, **40**, 3259–3264, https://doi.org/10.1002/grl.50636.

——, and ——, 2015: The contribution of atmospheric rivers to precipitation in Europe and the United States. *J. Hydrol.*, **522**, 382–390, https://doi.org/10.1016/j.jhydrol.2014.12.010.

Lawrence, D., 2016: Klimaendringer og framtidige flommer i Norge (in Norwegian). Norwegian Water Resources and Energy Directorate Rep. 81-2016, 66 pp., http://publikasjoner.nve.no/rapport/2016/rapport2016_81.pdf.

——, and H. Hisdal, 2011: Hydrological projections for floods in Norway under a future climate. Norwegian Water Resources and Energy Directorate Rep. 5-2011, 47 pp., http://publikasjoner.nve.no/report/2011/report2011_05.pdf.

Lussana, C., T. Saloranta, T. Skaugen, J. Magnusson, O. E. Tveito, and J. Andersen, 2018: seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the present day. *Earth Syst. Sci. Data*, **10**, 235–249, https://doi.org/10.5194/essd-10-235-2018.

Maraun, D., 2016: Bias correcting climate change simulations - A critical review. *Curr. Climate Change Rep.*, **2**, 211–220, https://doi.org/10.1007/s40641-016-0050-x.

Müller, M., and Coauthors, 2017: AROME-MetCoOp: A Nordic convective-scale operational weather prediction model. *Wea. Forecasting*, **32**, 609–627, https://doi.org/10.1175/WAF-D-16-0099.1.

Nash, J. E., and J. V. Sutcliffe, 1970: River flow forecasting through conceptual models. Part I - A discussion of principles. *J. Hydrol.*, **10**, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6.

Neiman, P. J., A. B. White, F. M. Ralph, D. J. Gottas, and S. I. Gutman, 2009: A water vapor flux tool for precipitation forecasting. *Water Manage.*, **162**, 83–94, https://doi.org/10.1680/WAMA.2009.162.2.83.

Olsson, J., and Coauthors, 2016: Hydrological climate change impact assessment at small and large scales: Key messages from recent progress in Sweden. *Climate*, **4**, 39, https://doi.org/10.3390/cli4030039.

Orvedal, K., and I. O. Peereboom, 2014: Flaumsonekart Delprosjekt Førde (in Norwegian). Norwegian Water Resources and Energy Directorate Rep. 61/2014, 42 pp., http://publikasjoner.nve.no/rapport/2014/rapport2014_61.pdf.

Persson, A., 2015: User guide to ECMWF forecast products. E. Andersson and I. Tsonevsky, Eds., ECMWF, 129 pp.

Porter, J. J., and S. Dessai, 2017: Mini-me: Why do climate scientists' misunderstand users and their needs? *Environ. Sci. Policy*, **77**, 9–14, https://doi.org/10.1016/j.envsci.2017.07.004.

Prein, A. F., and Coauthors, 2015: A review of regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53**, 323–361, https://doi.org/10.1002/2014RG000475.

Ralph, F. M., and M. D. Dettinger, 2011: Storms, floods, and the science of atmospheric rivers. *Eos, Trans. Amer. Geophys. Union*, **92**, 265–266, https://doi.org/10.1029/2011EO320001.

——, and Coauthors, 2017: Atmospheric rivers emerge as a global science and applications focus. *Bull. Amer. Meteor. Soc.*, **98**, 1969–1973, https://doi.org/10.1175/BAMS-D-16-0262.1.

Roald, L. A., 2008: Rainfall floods and weather patterns. NVE Consultancy Rep. 14-2018, 44 pp., http://publikasjoner.nve.no/oppdragsrapportA/2008/oppdragsrapportA2008_14.pdf.

Sælthun, N. R., 1996: The Nordic HBV model. Description and documentation of the model version developed Tech. Rep. NVE-E-PUB-07/96, Norwegian Water Resources and Energy Administration Publ. 7, 27 pp.

Sandvik, M. I., A. Sorteberg, and R. Rasmussen, 2018: Sensitivity of historical orographically enhanced extreme precipitation events to idealized temperature perturbations. *Climate Dyn.*, **50**, 143–157, https://doi.org/10.1007/s00382-017-3593-1.

Sharma, A., C. Wasko, and D. P. Lettenmaier, 2018: If precipitation extremes are increasing, why aren't floods? *Water Resour. Res.*, **54**, 8545–8551, https://doi.org/10.1029/2018WR023749.

Shepherd, T. G., 2019: Storyline approach to the construction of regional climate change information. *Proc. Roy. Soc.*, **475A**, 20190013, https://doi.org/10.1098/rspa.2019.0013.

——, and Coauthors, 2018: Storylines: An alternative approach to representing uncertainty in physical aspects of climate change. *Climatic Change*, **151**, 555–571, https://doi.org/10.1007/s10584-018-2317-9.

Sillmann, J., V. V. Kharin, F. W. Zwiers, X. Zhang, and D. Bronaugh, 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part II. Future climate projections. *J. Geophys. Res. Atmos.*, **118**, 2473–2493, https://doi.org/10.1002/JGRD.50188.

Sorteberg, A., D. Lawrence, A. V. Dyrrdal, S. Mayer, and K. Engeland, Eds., 2018: Climate changes in short duration extreme precipitation and rapid onset flooding – Implications for design values. NCCS Rep 1/2018, 143 pp., https://cms.met.no/site/2/klimaservicesenteret/rapporter-og-publikasjoner/_attachment/13537?_ts=163df95ff7b.

Sterl, A., and Coauthors, 2008: When can we expect extremely high surface temperatures? *Geophys. Res. Lett.*, **35**, L14703, https://doi.org/10.1029/2008GL034071.

Stohl, A., Forster, C., and 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. *J. Geophys. Res.*, **113**, D05102, https://doi.org/10.1029/2007JD009006.

Tveito, O. E., 2002: Spatial distribution of winter temperatures in Norway related to topography and large-scale atmospheric circulation. *IAHS Publ.*, **309**, 186–194.

——, I. Bjørdal, A. O. Skjelvåg, and B. Aune, 2005: A GIS-based agro-ecological decision system based on gridded climatology. *Meteor. Appl.*, **12**, 57–68, https://doi.org/10.1017/S1350482705001490.

van der Linden, E. C., R. J. Haarsma, and G. van der Schrier, 2018: Resolution-dependence of future European soil moisture droughts. *Hydrol. Earth Syst. Sci. Discuss.*, https://doi.org/10.5194/hess-2018-226.

van Haren, R., R. J. Haarsma, H. De Vries, G. J. Van Oldenborgh, and W. Hazeleger, 2015: Resolution dependence of circulation forced future central European summer drying. *Environ. Res. Lett.*, **10**, 055002, https://doi.org/10.1088/1748-9326/10/5/055002.

Vannforeningen, 2018: Drought and floods in 2018 – Is this climate change? (in Norwegian). Vannforeningen, https://vannforeningen.no/torke-og-%ef%ac%82om-i-2018-er-det-klimaendringer/.

van Vuuren, D. P., and Coauthors, 2011: The representative concentration pathways: An overview. *Climatic Change*, **109**, 5, https://doi.org/10.1007/S10584-011-0148-Z.

Vormoor, K., D. Lawrence, L. Schlichting, D. Wilson, and W. K. Wong, 2016: Evidence for changes in the magnitude and frequency of 70 observed rainfall vs. snowmelt driven floods in Norway. *J. Hydrol.*, **538**, 33–48, https://doi.org/10.1016/j.jhydrol.2016.03.066.

Whan, K., J. Sillmann, N. Schaller, and R. Haarsma, 2020: Future changes in atmospheric rivers and extreme precipitation in Norway. *Climate Dyn.*, **54**, 2071–2084, https://doi.org/10.1007/s00382-019-05099-z.

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

Paper IV:

# The role of spatial and temporal model resolution in a flood event storyline approach in western Norway

**N. Schaller, J. Sillmann, M. Müller, R. Haarsma, W. Hazeleger, T. J. Hegdahl, T. Kelder, G. van den Oord, A. Weerts, and K. Whan**
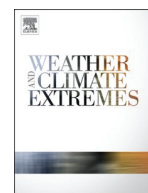
IV

# The role of spatial and temporal model resolution in a flood event storyline approach in western Norway

Nathalie Schaller [a,*], Jana Sillmann [a], Malte Müller [b], Reindert Haarsma [c], Wilco Hazeleger [d], Trine Jahr Hegdahl [e], Timo Kelder [f], Gijs van den Oord [g], Albrecht Weerts [h,i], Kirien Whan [c]

[a] Center for International Climate Research (CICERO), Gaustadalleen 21, 0349 Oslo, Norway
[b] The Norwegian Meteorological Institute, Postboks 43 Blindern, 0371 Oslo, Norway
[c] Royal Netherlands Meteorological Institute (KNMI), De Bilt, the Netherlands
[d] Faculty of Geosciences, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, the Netherlands
[e] Norwegian Water Resources and Energy Directorate (NVE), Middelthuns Gate 29, 0368 Oslo, Norway
[f] Geography and Environment, Loughborough University, Loughborough, UK
[g] Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, the Netherlands
[h] Deltares, Postbus 177, 2600 MH Delft, the Netherlands
[i] Hydrology and Quantitative Water Management Group, Wageningen University, Wageningen, the Netherlands

## ARTICLE INFO

## ABSTRACT

We apply a physical climate storyline approach to an autumn flood event in the West Coast of Norway caused by an atmospheric river to demonstrate the value and challenges of higher spatial and temporal resolution in simulating flood impacts. We use a modelling chain whose outputs are familiar and used operationally, for example to issue flood warnings. With two different versions of a hydrological model, we show that (1) the higher spatial resolution between the global and regional climate model is necessary to realistically simulate the high spatial variability of precipitation in this mountainous region and (2) only with hourly data are we able to capture the fast flood-generating processes leading to the peak streamflow. The higher resolution regional atmospheric model captures the fact that with the passage of an atmospheric river, some valleys receive high amounts of precipitation and others not, while the coarser resolution global model shows uniform precipitation in the whole region. Translating the event into the future leads to similar results: while in some catchments, a future flood might be much larger than a present one, in others no event occurs as the atmospheric river simply does not hit that catchment. The use of an operational flood warning system for future events is expected to facilitate stakeholder engagement.

## 1. Introduction

Between the 27th and 29th of October 2014, large amounts of precipitation - up to 300 mm in less than 5 days in some areas - fell over the West Coast of Norway, causing floods and damages in several valleys (Lussana et al., 2018). The towns of Flåm and Odda were particularly affected, with bridges destroyed and houses dragged into the river. The West Coast of Norway is the wettest region in Europe (e.g. Lussana et al., 2018) and due to its steep topography (see SI Fig. 1), often experiences floods. Spring floods are usually associated with snow melt, summer floods with convective precipitation and autumn floods with atmospheric rivers (Dyrrdal et al., 2016). The large amounts of rain in October 2014 were caused by the passage of such an atmospheric river

(Lussana et al., 2018 and SI Fig. 2a). Atmospheric rivers are long and narrow regions of intense water vapour transport in the lower atmosphere and Sodemann and Stohl (2013) and Azad and Sorteberg (2017) have shown how important they are to transport sub- and extratropical moisture to the West Coast of Norway. Benedict et al. (2019a) further showed that atmospheric rivers are associated with more than 85% of extreme precipitation in this region for the period September–March from 1979 to 2014.

But even in a region used to heavy precipitation and floods, it is a challenge to decide what (and to what extent) adaptation measures should be considered for the future, when even more precipitation is projected by climate models (Whan et al., 2019). Making sense of and using probabilistic climate projections for extreme weather events is

---

challenging for most decision makers and end-users (Porter and Dessai, 2017). One of the main issues is the coarse resolution, both in a space and time, provided by Global Coupled Models and Regional Coupled Models (GCMs and RCMs), which is often not adapted to the user needs. Another difficulty lies in the interpretation of the model output, especially the often large ranges of outcomes obtained when combining different GCMs and RCMs, while users seem often to prefer a single value with a probability attached to it. Probabilistic climate projections are certainly the basis of our knowledge on future climate change and its uncertainties, but there is a growing demand for climate change information that is more relevant, useful and tailored to user needs, especially on the topic of climate impacts (Shepherd et al., 2018). Bottom-up approaches, where a study is designed with users and their needs discussed upfront, rather than climate scientists assuming what information is useful, are to be favored (Dee et al., 2011).

Storyline approaches, where "tales" about possible future weather events are created, are currently being advanced as a way to complement information from the more classic probabilistic assessments and to better feed the imagination of users (Hazeleger et al., 2015). Physical climate storylines are physically self-consistent unfolding of past events, or of plausible future events, which are conditioned on a set of assumptions and built from causal arguments, but mainly inspired by observed or anticipated high impact events. They are particularly designed to improve risk awareness, to strengthen decision-making (e.g. proactive adaptation), to explore the boundaries of plausibility, and to provide a physical basis for partitioning uncertainty (Shepherd, 2019).

Since in a physical climate storyline study, the focus is on specific events with limited spatial and temporal scope, the computational requirements are smaller in terms of numbers of time steps, climate models and ensemble members used, and therefore higher resolution versions of a climate model can be more easily run. However, one needs to decide how to constrain the simulations for a specific event. The assumption is that by having a higher resolution in both time and space, and ideally, data formats already known by the user, or used operationally, will provide more relevant information for users as this level of detail cannot be achieved currently in a full probabilistic context given computing limitations (Hegdahl et al., 2020).

A key aspect of a physical storyline is therefore the modelling chain used. Recently, studies have reported successful coupling of several models beyond the established GCM-RCM-hydrological model chain. For example, Felder et al. (2018) demonstrated the feasibility of a full and comprehensive model chain from the atmospheric to local scale flood loss models. In the context of probabilistic event attribution, Schaller et al. (2016) went from GCM simulations of the particular weather situation during the 2013/2014 winter over the North Atlantic/British Isles to the count of properties at risk of flooding in the Thames catchment. While probabilistic event attribution compares extreme events in the present versus the same extreme event in an alternative present where there would have been no human influence, physical climate storyline approaches attempt to translate an extreme event that has happened and is of interest to users into the future, where some level of human activity is assumed in terms of anthropogenic GHG emissions and/or socio-economic changes (e.g. different management practices) (Shepherd et al., 2018; Sillmann et al., 2019). Both research fields, probabilistic event attribution and physical climate storylines, could benefit from the other by lessons learned as they develop. For example, in the case of flood events, it is key to at least use one hydrological model rather than stopping at the heavy precipitation event. Each river catchment has very specific properties that cannot be generalized, and in some cases, even if a signal is found from the GCMs and RCMs, it could be diluted at the hydrological model step, as in Schaller et al. (2016).

The overall aim of this study is to investigate how a high-impact extreme event in the future would look like compared to one in the present climate, using a physical climate storyline approach and an operational modelling chain.

As mentioned above, there is an assumption that physical climate storyline approaches, with their higher spatial and temporal resolutions, provide an added value compared to more standard probabilistic multimodel/multi-scenario approaches such as, e.g., the Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al., 2012) or the European Coordinated Regional Downscaling Experiment (EURO--CORDEX, Jacob et al., 2014) (Hazeleger et al., 2015). This seems particularly relevant in the case of floods associated with extreme precipitation in a mountainous region such as the West Coast of Norway, as indicted in e.g. Müller et al., 2017; Prein, 2015. In addition, climate information can best be used by users when it is provided in a format they are familiar with. In this study, we are therefore using the operational modelling chain for flood forecasting in Norway to show the effect of having higher temporal and spatial resolution models on the simulated streamflows.

Section 2 describes the methods used for the event selection and model simulations performed. In Section 3, we compare the simulated and actual events, and put the differences between the present and future events into the context of global warming. Then we present and discuss the effect of higher spatial and temporal model resolution to simulate streamflows in specific catchments, and present conclusions in Section 4.

## 2. Methods

### 2.1. Stakeholder involvement and event selection

The Norwegian Meteorological Institute and Statkraft (the largest energy provider in Norway) are stakeholders directly involved in the project and provided insights for the event selection. During the kick-off meeting of the project at the end of 2016, different extreme events that occurred in Norway in the recent past were further discussed with a larger group involving further stakeholders, such as municipalities, television channels and state authorities (https://www.cicero.oslo.no/en/posts/single/making-sense-of-future-climate). The events that sparked most interest were the September 2005 and October 2014 floods hitting western Norway, and particularly causing severe damages in the city of Bergen and the touristic fjord village Flåm, respectively (Iden et al., 2005; Lussana et al., 2018; Stohl et al., 2008). For this reason, the analogue event considered in this study is an October flood caused by an atmospheric river (AR).

The premise of this study is to use an existing and operational modelling chain to ease the use of the results to the different potential users. The Norwegian Meteorological Institute uses the 6-hourly short range weather forecast from the ECMWF model, downscales them using the AROME-MetCoOp model for Scandinavia, and the output is used by the state meteorologists to issue weather forecasts, but also by the Norwegian Water Resources and Energy Directorate (NVE), to issue flood warnings. We run two different hydrological models that use different temporal and spatial input from the weather forecast to show the added value for the users of having high resolution climate and hydrological models. The rest of this section describes in more details the different types of models (i.e. global climate model, regional weather forecasting model and hydrological models) used to reproduce as closely as possible the operational setup for weather forecasts and flood warnings in Norway.

### 2.2. The EC-Earth high-resolution global climate model and AR events

#### 2.2.1. EC-Earth control simulations

The global climate model used is EC-Earth v2.3 model (Hazeleger et al., 2010) at a resolution of T799 L91 (~25 km), which was the operational resolution at the European Centre for Medium Range Weather Forecasts (ECMWF). This spatial resolution is much higher than in any of the CMIP5 models (Taylor et al., 2012) and is able to resolve tropical cyclones (Haarsma et al., 2013). Different model runs were

performed for the present (2002–2006) and future (2094–2098) climate, which are referred to as EC-Earth PRESENT and FUTURE in the rest of the text. Each of these datasets consists of a 6-member ensemble spanning 5 years resulting in a 30-year dataset. In present simulations, observed greenhouse gas and aerosol concentrations were applied while future concentrations were derived from the RCP 4.5 scenario (van Vuuren et al., 2011). Only the atmosphere and land surface (HTESSEL; van den Hurk et al., 2000) are solved explicitly by the model to allow for the generation of high-resolution results spanning an extensive period of time. Therefore, sea surface temperatures (SSTs) were imposed using daily data at 0.25° resolution from NASA (http://www.ncdc.noaa. gov/oa/climate/research/sst/oi-daily.php) for the 2002–2006 period. An estimate of the SST for the other periods was made by adding the ensemble mean predicted change using ECHAM5/MPI-OM in the ESSENCE project (Sterl et al., 2008) which used the SRES A1B scenario. This is comparable to RCP 4.5 scenario, having a slightly larger global temperature increase by the end of the twenty first century (Rogelj et al., 2012; Haarsma et al., 2013). The 5 independent ensemble members were generated by small perturbations of the atmospheric initial conditions; for further details on this model setup and spin-up we refer to Haarsma et al. (2013) and Benedict et al. (2019b). The output data was stored on 5 pressure levels (850, 700, 500, 300 and 200 hPa) at 6-hourly intervals, surface fields were saved on 3-hourly basis. For each member a restart state was saved on the first of January of every simulated year, ensuring bit-exact reproducibility of the generated weather at the computing facility at KNMI.

### 2.2.2. Selection of AR events in EC-Earth in present and future climate

The sum of the large scale and convective precipitation fields produced by EC-Earth served as the basis for the analysis of extreme rainfall events. The daily cumulant of this quantity within the West Coast region provided the distribution of which the days above 99.98 percentile were selected. This translates to a lower bound of 77 mm/day for the present day and 89 mm/day for the future climate datasets. This percentile was chosen to yield two events for each period. Of the two events in the present, one occurred in October and the other in November, and in the future, one occurred in October and one in January. To avoid potential inconsistencies with precipitation falling as snow in November and January, we chose only the October events in both present and future.

A brief model evaluation of these EC-Earth simulations for the month of October is presented in SI Fig. 2, since this is the month in which the selected events occur. Haarsma et al. (2013) however provide a more formal evaluation of the model. Whan et al. (2019) further found that EC-Earth has a good representation of the annual cycle of AR frequency compared to ERA-Interim (Dee et al., 2011), with most events occurring in September to January. The number of AR events per year in ERA-Interim and EC-Earth is also very similar.

### 2.3.3. Perturbation of AR events in EC-Earth

After the event selection described above, we have used the same EC-Earth binary to generate 10-member 'forecast' ensembles for each of these events (referred to as EC-Earth reruns in the following). From the existing restart states, we ran the model until 5 days before the precipitation peak and let it write a new restart state. Because we used the exact restarts on the same machine, we were able to identically recreate the selected events. Then, from these snapshots 10 ensemble members are created using stochastically perturbed parametrization tendencies (SPPT) in the Integrated Forecasting System. This method adds multiplicative noise to the model physics tendencies, representing the arbitrariness of the model's subgrid parametrization schemes:

$$X_{tot} = X_{dyn} + (1+\rho)X_{phys}$$

where $X_{tot}$ denotes the total tendency, $X_{dyn}$ the tendency from explicit dynamics and $X_{phys}$ arises from vertical model physics such as radiation, clouds, convection and turbulence. All prognostic variables (wind

components, temperature and humidity) are subject to this randomization and $\rho$ are random numbers on a latitude-longitude grid of $2.\overset{\circ}{5}$ x $2.\overset{\circ}{5}$ which are re-generated every 5 time steps, i.e. 50 min. Within each of these grid boxes, the multiplicative noise on the physics tendencies is uniformly distributed within the interval $-0.5 < \rho < 0.5$. For these members we used the same SST and aerosol and greenhouse gas forcings as for the original datasets. The procedure results in an ensemble of 10 global runs which all give rise to atmospheric rivers and high precipitation rates at the selected event dates. These members were run for 7 days to capture the full event and all prognostic fields were saved on all model levels for downscaling purposes. Because we chose a moderate perturbation of the tendencies and started only 5 days before the occurrence of the extreme event, the members are correlated and display all high precipitation rates at the event date.

### 2.3. The regional operational weather forecast model AROME-MetCoOP

The regional weather prediction system AROME-MetCoOp (Müller et al., 2017) used for the operational weather forecast in Norway, is used here to downscale the atmospheric simulations from EC-Earth. The AROME-MetCoOp model covers large parts of the Nordic countries with a horizontal resolution of 2.5 km, on a 750 by 960 grid, and with 65 layers in the vertical where the lowest level is at 12 m. The model has non-hydrostatic dynamics, semi-Lagrangian advection and two-time level semi-implicit time stepping using a 75 s time step. Parameterisation of physical processes are described in Bengtsson et al. (2017). AROME-MetCoOp is a particular configuration of the HARMONIE system suited for the highest resolutions. HARMONIE (Hirlam Aladin Regional/Meso-scale Operational NWP In Europe) is a cooperation including Meteo-France and their Aladin partners, the Hirlam group and also ECMWF with their IFS (Integrated Forecasting System) model. The configuration used in this study (harmonie-40h1.1.rc.1) is very close to the operational weather forecasting version used in MetCoOp in early 2018.

The AROME-MetCoOp operational weather prediction system (referred to as AROME in the following) has been evaluated in detail in Müller et al. (2017) and they showed that forecasting of precipitation is strongly improved compared to the global ECMWF weather prediction system (horizontal resolution T1279, approx. 16 km), which is used as input for AROME in the operational weather forecast for Norway. Especially, during extreme events of large-scale precipitation, the magnitudes and locations of maximum precipitation are more consistent with observations.

For this study, AROME is forced by 3-hourly EC-Earth simulations at the lateral and upper boundaries. The model is initialized by the EC-Earth fields 36 h before the extreme event and forecasts up to a lead time of 144 h (6 days) are performed.

### 2.4. The operational flood forecasting model HBVlump

The Hydrologiska Byråns Vattenbalans (HBV) model (Bergström, 1976) as described in Sælthun (1996) and Beldring (2008) is used by the operational flood forecasting service in Norway. The operational HBV model (in the following referred to as HBVlump) uses daily catchment average values of temperature and precipitation as input. The upstream area of a streamflow measuring point defines the catchment area. Each catchment is divided into ten equally large elevation zones, and the average temperature and precipitation is elevation adjusted to each elevation zone. The model consists of a vertical structure with a soil moisture routine, a snow routine, and response functions for quick and slow runoff.

The HBVlump is calibrated for the 1996–2012 period, and thereafter validated for the 1980–1995 period, using temperature and precipitation from the seNorge v1.1 dataset (Mohr, 2008, 2009) as observed input, and the measured streamflow from the NVE database as a reference. Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) and volume

bias are the chosen calibration metrics. In this study, we focus on two catchments, Flåm, which flooded in October 2014, and Røykenes, a catchment representative for the September 2005 flood that affected the city of Bergen. Røykenes has a Nash-Sutcliffe efficiency of 0.82 for the calibration period and 0.81 for the validation, with a volume bias of 10%, whereas Flåm has Nash-Sutcliffe efficiency of 0.82 for the calibration and 0.71 for the validation, with volume bias of 7%..

## 2.5. The hydrological model HBVdist

A platform for distributed hydrological modelling, wflow_hbv, is used to set up the distributed HBV model (Schellekens et al., 2017). This version of the model, using gridded input data rather than catchment-averaged values, can incorporate geographical information (e.g. elevation, geology and land-use) and solves the water balance for each grid cell. The water balance is solved in the same way as previously described for HBVlump and the resulting quick and slow runoff are forced into a kinematic wave routing model. Pre-processing of the input data is performed in Delft-FEWS (Werner et al., 2013).

The distributed HBV model is calibrated on $1 \times 1$ km spatial and hourly time resolution. A gridded hourly version of the seNorge v.2.0 dataset (seNorge2 in the following, Lussana et al., 2018) is used to provide observed temperature and precipitation input and hourly streamflow measurements are used as reference. Because the hourly temperature and precipitation series are much shorter than the daily version, the calibration was performed over the period 2010–2013 and the years 2014–2016 were used for validation. No streamflow observations were available for the year 2015 in Flåm because the measurement station was destroyed during the flood in 2014, meaning that there are only two years of data for model validation for this catchment. In consultation with NVE, the parameters representing field capacity ('FC'), snowmelt ('TT'), snow/rainfall partitioning ('TTI'), seepage through the soil layer ('BetaSeepage'), and recession coefficients influencing the slow and quick runoff ('K4' and 'KHQ') were selected as most representative of the dominant characteristics in the Norwegian catchments. These parameters were selected for calibration with similar ranges for the parameters as previously reported in Nordic studies (Swedish Meteorological and Hydrological Institute, 2014).. The python package for parameter estimation (SPOTPY) is used to perform the calibration, with 400 Monte Carlo samples, assuming uniform distribution of the parameters and a spin-up period of two months (Houska et al., 2015). Because of its focus on peaks, Nash-Sutcliffe efficiency is used as objective to compare simulations with observations (Nash and Sutcliffe, 1970). For Flåm, the Nash-Sutcliffe efficiency is 0.67 for the calibration phase and 0.69 the validation period. For Røykenes, the Nash-Sutcliffe efficiency is 0.82 for both the calibration the validation phase.

## 2.6. Initial hydrological conditions

The HBV model has storages for groundwater, soil-moisture and snow, where especially snow and soil-moisture are important for the development of floods in Norwegian catchments. Since streamflow response depends on the water storage within a catchment, there is consequently not a unique relationship between precipitation intensities and flood sizes (Beldring et al., 2008). To spin up the hydrological models and thereby define the hydrological states the models are run using the seNorge2 interpolated temperature and precipitation observations. The states at the end of the spin-up period are used as initial hydrological conditions for all present and future event simulations.

HBVlump was run with daily observations 1 year prior to 25th October 2014. The states from October 2014 are chosen for two reasons: in October 2014 there were a situation with AR induced floods (Lussana et al., 2018) and in addition, the weather in October 2014 was wet and mild (small snow storages), and therefore similar to what is expected to occur in a future climate in western Norway (e.g. Hanssen-Bauer et al.,

2017).

In the HBVdist setup, ten initial conditions are sampled to incorporate variability arising from plausible autumn initial hydrological conditions other than the 2014 event. The HBVdist model was run over the years 2010–2014 and the hydrological model states are selected on the 10th of September and 10th of October. Note that in this approach the same initial conditions are used for the present and future climate, whereas a reduction in the snow reservoir in the future is very likely (Hanssen-Bauer et al., 2017). However, in autumn in Norway, extreme precipitation is the dominant driver of floods (Beldring, 2008) and the amount of snow is still building up and is relatively small, so therefore the overestimation of future initial snow conditions is assumed to be of little importance for the flood magnitude.

## 3. Results and discussion

### 3.1. The event - as observed and simulated

The event on the 27th and 28th of October 2014 in Flåm, is shown in SI Fig. 2 using the ERA-Interim reanalysis in terms of integrated water vapour transport (IVT). ARs are objectively identified by applying a detection algorithm based on the IVT field, as described in Whan et al. (2019). The two selected modelled events are shown in SI Fig. 2 b-c. 10 reruns are then performed for each of these two present and future events. As expected, all the reruns also have an AR, however, the location and intensity in terms of IVT varies. SI Fig. 3 further shows that EC-Earth has a small positive bias in terms of IVT over the North Atlantic Ocean in general compared to ERA-Interim but is generally able to represent ARs statistics properly (Whan et al., 2019).

SI Fig. 3 shows that the October climatology in EC-Earth reproduces well different percentiles of daily accumulated precipitation along the West Coast region compared to seNorge2. Despite having different horizontal resolutions (25 km and 2.5 km), both EC-Earth and AROME (see Müller et al., 2017) perform well in terms of precipitation in the West Coast of Norway and are therefore appropriate models to perform this storyline study.

The observed accumulated precipitation in the West Coast region (defined as 60º-61ºN, 5.2º-7.5ºE) over a 6-day period, from 25th until 30th of October 2014, using seNorge2 is shown in Fig. 1. In the present, the 10 ensemble members from EC-Earth reruns and AROME show similar accumulated precipitation amounts, although the time evolution of the events differs slightly, especially in the first 30 h. The reason for this is that the October 2014 event was characterised by two peaks in precipitation, whereas the events selected in the EC-Earth simulations had a single peak.

Fig. 1 further shows how a future heavy precipitation event also caused by an AR in October could look like in the West Coast region. Overall, the accumulated precipitation after 6 days is larger in these future events compared to the present events, with a few ensemble members in both the EC-Earth reruns and AROME having similar or lower amounts compared to the 2014 event. Especially in the present events, EC-Earth appears wetter compared to AROME. SI Figs. 5 and 6 show maps of 6 days accumulated precipitation for the present and future events in all ensemble members. From these maps, it is clear that AROME does simulate higher local peaks, but also many areas in-between with very little precipitation, whereas EC-Earth is wet in every grid cell. This is likely the reason why, when looking at a larger area such as the West Coast region defined here, a coarser resolution model might appear wetter. This result shows that using a high-resolution regional atmospheric model such as AROME is beneficial to give a more realistic and more complete picture of the chance of a catchment being hit. Using a high-resolution GCM will indicate that the passage of an atmospheric river leads to precipitation in each grid cell, which is not according to observations. The localised nature of precipitation is something that users experience every day, so using a higher resolution regional model that reproduces this characteristic should give
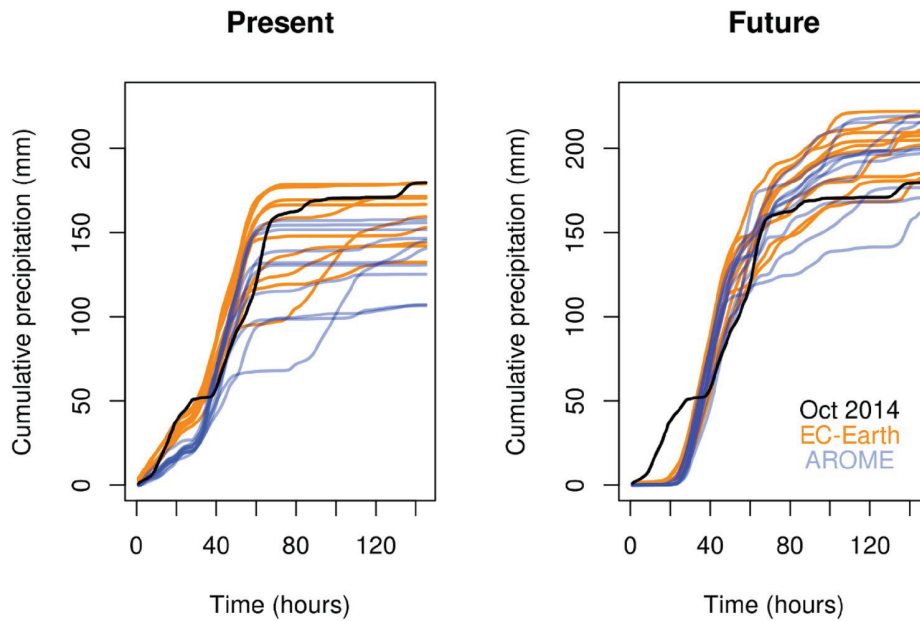
## Present

## Future



**Fig. 1.** Cumulative distribution of precipitation during the October 2014 event in seNorge2 (black) and each of the ten ensemble members of AROME (blue) and EC-Earth (orange) in the present (left) and future (right), for the region in 60°-61°N, 5.2°-7.5°E. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

more trust in the story.

To put these results in context, extreme and mean precipitation are expected to increase both during the cold and warm season in Northern Europe according to the CMIP5 models (Hodnebrog et al., 2019; Sillmann et al., 2013). However, when using a higher resolution model (WRF driven by CESM1-CAM4), Hodnebrog et al. (2019) finds that June–September mean precipitation decreases by the end of the century (RCP8.5 scenario) in south-east Norway, but slightly increase along the West Coast of Southern Norway. EC-Earth shows a very similar pattern for October precipitation changes (see Fig. 2a, changes in 2 m temperature and IVT are shown in SI Fig. 7). Understanding precisely what drives this future precipitation change pattern is beyond the scope of this article, but we briefly attempt to quantify whether there are some changes in dynamics that could explain this drying in particular, since the CMIP5 models, and the Clausius-Clapeyron relationship, hint rather

to a wettening (Hodnebrog et al., 2019).

Using the Weather Types (WT) classification from Otero et al. (2018), we quantify the changes in frequency of these WT for October in EC-Earth between present and future simulations. Haarsma et al. (2013) however provide a more formal evaluation of the model.

SI Fig. 8 shows that there are indeed some significant changes in WT occurring in our region of interest. The increase in Westerly flow is in line with an increase in precipitation along the northern side of the West Coast (perhaps strengthening the wettening expected from thermodynamics). Interestingly, there seems to be a strong increase in anticyclonic conditions in southern Norway at the same time, which is consistent with the drying in this region. The strong decrease in cyclonic conditions is more difficult to relate to the increase in precipitation in the northern part of the West Coast. However, precipitation does not only occur under cyclonic conditions hence this result is not necessarily
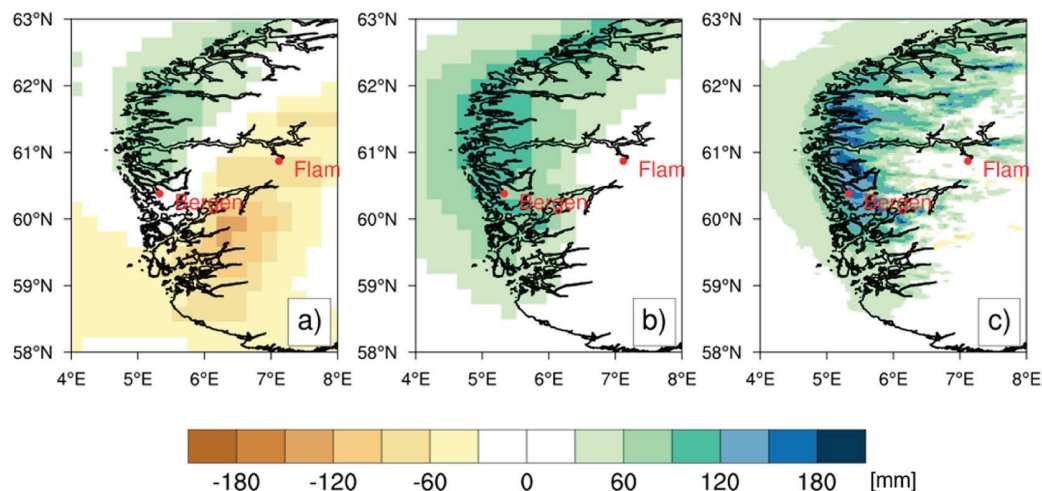


**Fig. 2.** Maps of the difference in a) 30 years October precipitation averages between PRESENT and FUTURE EC-Earth simulations, b) 144 h accumulated precipitation averaged over 10 ensemble members of the present and future EC-Earth reruns of the event and c) 144 h accumulated precipitation averaged over 10 ensemble members of the present and future AROME simulations of the event.

contradicting the other findings and more research would be needed to provide a proper interpretation. The results of the Westerly flow and Cyclonic conditions weather types are in line with Otero et al., 2018 (Fig. 5 in Otero et al., 2018) shows a similar pattern for SON in CMIP5), although the increase in Anticyclonic conditions is less pronounced in the southern part of Norway in CMIP5, potentially explaining why these models do not show a drying there.

The future precipitation event in this storyline study is wetter compared to the present one in both the EC-Earth reruns and AROME simulations over the western and northern parts of the West Coast of Norway, as shown in Fig. 2b and c. This is in line with both the facts that the intensity and frequency of ARs reaching the West Coast of Norway are expected to increase in the EC-Earth FUTURE simulations (Whan et al., 2019) and that most extreme precipitation events in the cold season are associated with ARs (Benedict et al., 2019a). It is also in line with the fact that despite the overall response of mean (or median) precipitation in the region being drying for October (see Figs. 2 and SI Fig. 4a, the higher percentiles of the distribution (95th, 99th and 99.99th) show more precipitation everywhere in the region in the future (shown as transects between 60°-61°N in SI Fig. 3b–d). We therefore seem to have captured this general response of extreme precipitation to global warming in that region with the present and future events we selected, as well as the fact that generally less precipitation reaches inland in the future in the case of ARs (Sandvik et al., 2018).

### 3.2. Influence on the spatial resolution of the driving model

The Norwegian operational flood-forecasting model (HBVlump) is used first to evaluate the effect of meteorological input data with higher horizontal resolution, i.e. an expected added-value of using AROME compared to EC-Earth, for the estimation of floods in the non-managed catchments of Norway. In this study, we focus on the Røykenes and Flåm catchments as discussed with stakeholders (see SI Fig. 1). The two catchments are interesting because they show a different response to the present and future AR's.

While considering the larger West Coast region defined in the previous section, EC-Earth's events were wetter compared to AROME (see Fig. 1), but when focusing on smaller catchment scales, this does not always hold. Whereas the precipitation is represented as one value over the entire catchment when using EC-Earth as input, valleys within the catchment can be either hit or not hit by precipitation when using AROME as input, as illustrated in SI Figs. 5–6 (lower 10 panels). For the two catchments selected in this study, we find that the precipitation and the resulting peak streamflows appears usually wetter for AROME than for EC-Earth (see Figs. 3–4). Another aspect is the difference in catchment average temperature between AROME and EC-Earth. For the present event in Flåm (see SI Fig. 9), the temperature modelled by AROME is about 1 °C warmer compared to EC-Earth. Whether the modelled temperature is below or above the melting point, as is the case in the present event in Flåm, will strongly affect the streamflows. The small streamflow response to precipitation of day 5 and 6 in Røykenes (Fig. 3) is clearly due to the low temperature (SI Fig. 9). In addition, with the steep topography of the region, differences in modelled precipitation and temperature at different elevations complicates the total estimations of streamflow.

In Røykenes, the future event is larger than the present one (see Fig. 3), which is in line with what one would expect for an extreme precipitation event in this region in a future climate (see previous section). For Flåm, however, the future event contains less precipitation and lower streamflow peaks than the present event for both the EC-Earth and the AROME (see Fig. 4). The Flåm catchment is not severely hit by the selected events, whereas the case of Røykenes exemplifies how the near coastal areas experience the highest increase in precipitation, which is in line with Sandvik et al. (2018). An evaluation of historical AR events shows that the development of the events and the area most severely hit varies between the events (e.g. Iden et al., 2005; Lussana et al., 2018; Stohl et al., 2008). Hegdahl et al. (2020) applied the AR events of this study to multiple catchments in the West Coast region and show that the future AR events cause more intense floods in more catchments than the present events. The overall impact is therefore
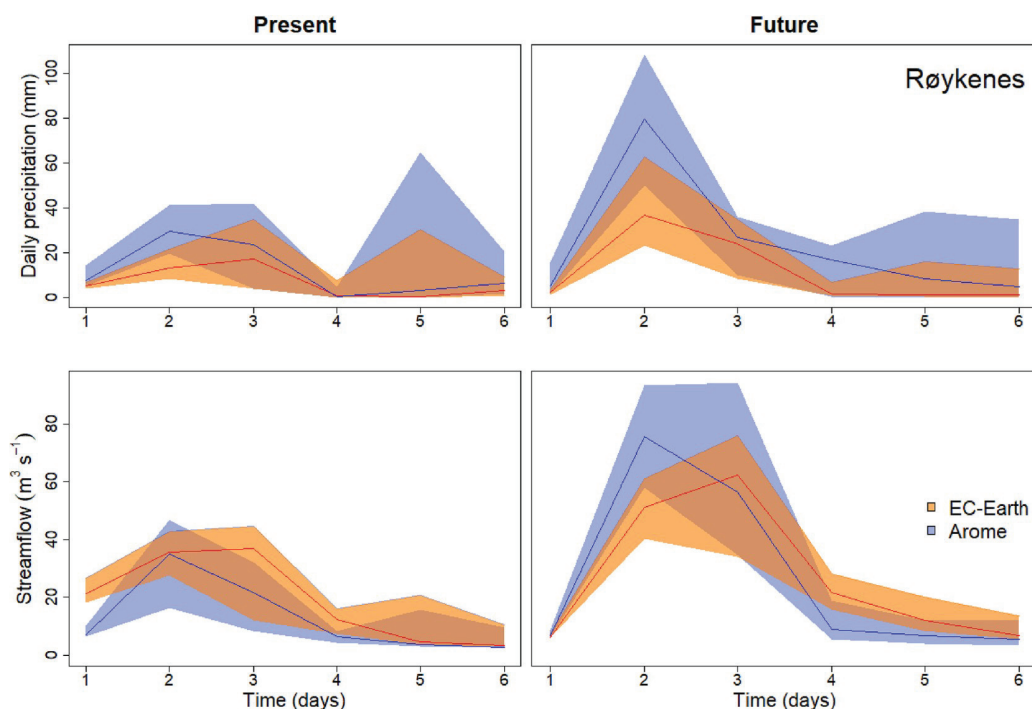


**Fig. 3.** Present (left) and future (right) events in the Røykenes catchment. The top panels show daily precipitation and the bottom ones, the streamflow. The shaded area represents the range of the 10 ensemble members for both EC-Earth (orange) and AROME (blue), with the lines within showing the median of the ten ensemble members. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
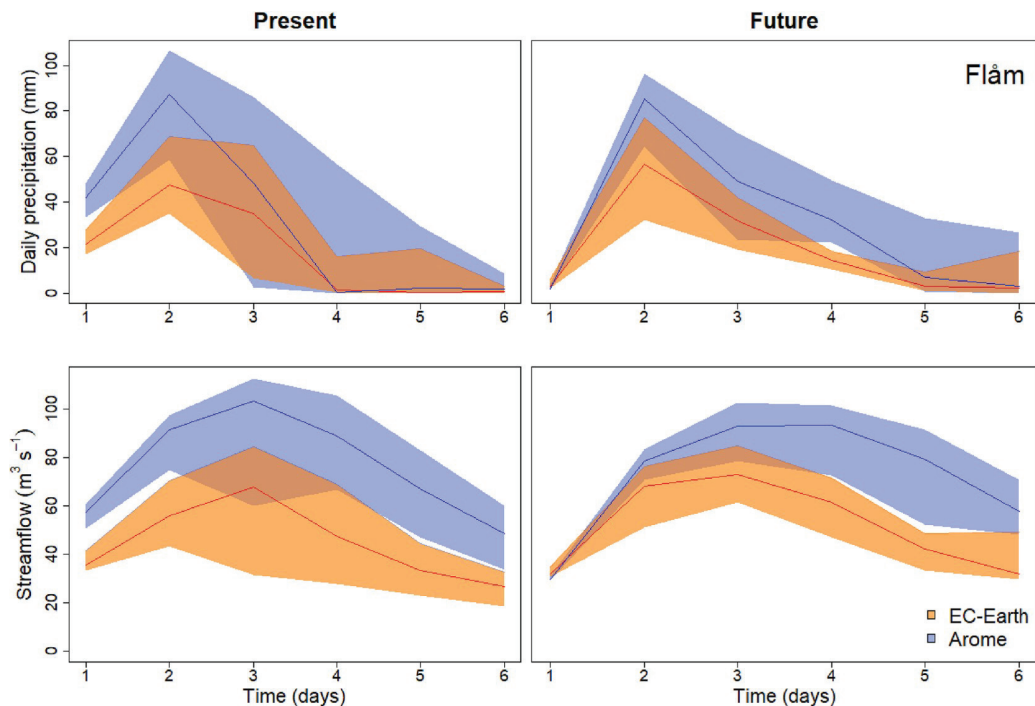
**Fig. 4.** Present (left) and future (right) events in the Flåm catchment. The top panels show daily precipitation and the bottom ones, the streamflow. The shaded area represents the range of the 10 ensemble members for both EC-Earth (orange) and AROME (blue), with the lines within showing the median of the ten ensemble members. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

higher for the future events than for present events, even though not all catchments are affected, as is the case for Flåm.

Figs. 3 and 4 further exemplify how combining AROME with a hydrological model is important to model the highest estimates of possible AR induced floods. AROME resolves the small-scale atmospheric processes, gives a better representation of orographic precipitation, and thereby provides an improved estimate for precipitation in relatively small and steep catchments, typical for the West Coast of Norway. The hydrological model describes the hydrological land process and estimates the hydrological initial conditions of the catchments that are important for the flood development. Besides precipitation, both snowmelt and soil moisture excess are drivers for floods (Beldring et al., 2008; Hegdahl et al., 2020; Sharma et al., 2018). Uncertainties related to the initial hydrological conditions are therefore discussed in the next section.

### 3.3. Influence of the higher temporal resolution

In this section, we are comparing streamflow simulations form the operational hydrological model set-up (HBVlump) to an alternative higher resolution distributed HBV model (HBVdist). HBVlump runs using daily temporal resolution, and precipitation and temperature data are averaged over the river catchment (268 $km^2$ for Flåm and 50 $km^2$ for Røykenes). HBVdist runs using hourly temporal resolution, and the input data are interpolated on a $1 \times 1$ km grid. A hydrological model running on a higher temporal resolution enables a better utilization of the available high-resolution temperature and precipitation data from AROME. From a theoretical perspective, the spatial distribution of precipitation is important because precipitation that falls in the upper areas of the catchment will take a longer time to contribute to the streamflow response than precipitation in the downstream area of the catchment. Similarly, the temporal resolution can influence the streamflow peak as higher intensity precipitation events can result in a quicker streamflow response.

Initial hydrologic conditions, such as the soil moisture or the snow

depth, are important factors that influence the streamflow peak (e.g. Bengtsson et al., 2017). To sample the hydrological uncertainty of the flood simulations, HBVdist is rerun with 10 different initial conditions. Additionally, to sample hydrological parameter uncertainty, 10 parameter sets are selected during the calibration phase. Note that the uncertainty ranges in Figs. 5 and 6 represent this hydrological uncertainty, while the uncertainty range in Figs. 3 and 4 represented the 10 different ensemble members of each simulation.

To test this model framework, we run the HBVdist for the October 2014 flood, forced with temperature and precipitation from seNorge2 observations and from AROME simulations (Fig. 5). Note that the AROME simulations used here are the actual archived forecasts for 2014 and not the AROME climate simulations downscaling EC-Earth. Fig. 5 shows that the measured peak streamflow in Flåm is within the range of the streamflows modelled by HBVdist using both seNorge2 and the AROME actual forecasts. For the Flåm catchment, the streamflow simulation using both seNorge2 and AROME are similar and the measured streamflow is within the simulated hydrological uncertainty range, giving confidence in the ability of the modelling chain used in this study to simulate flood events.

We can only present this validation for HBVdist for the Flåm catchment, because seNorge2 hourly data does not extend back to 2005 and AROME was not operational in 2005. Thus, it is not possible to produce a similar validation for the September 2005 event in Røykenes.

Fig. 6 shows the streamflow simulations of HBVdist and HBVlump for Flåm and Røykenes. The same AROME temperature and precipitation input data is used for both hydrological models. For clarity, only the AROME ensemble member that causes the highest streamflow simulations is shown in Fig. 6, while again the range represents the uncertainty of the initial conditions and hydrological model parameters.

For the observed 2014 event, we find that the flood was building up over three days, with the highest peak generated over 5 h: from 12 to 17pm on October 28th. To realistically describe the responses of the catchments, hydrological models on sub-daily resolution are required. As expected, HBVdist simulates higher streamflow peaks than HBVlump
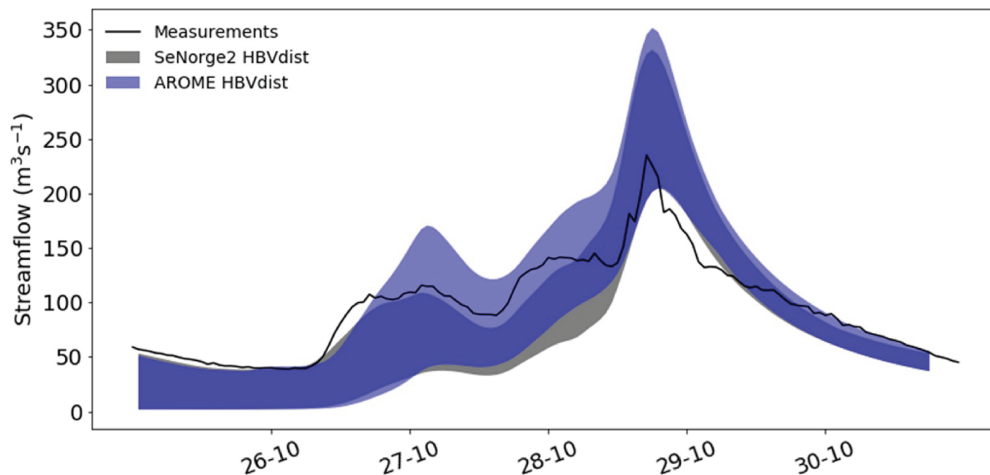
**Fig. 5.** Measured and HBVdist simulated streamflow for Flåm. The uncertainty range represents hydrologic uncertainty for the ensemble member with the highest peak streamflow.
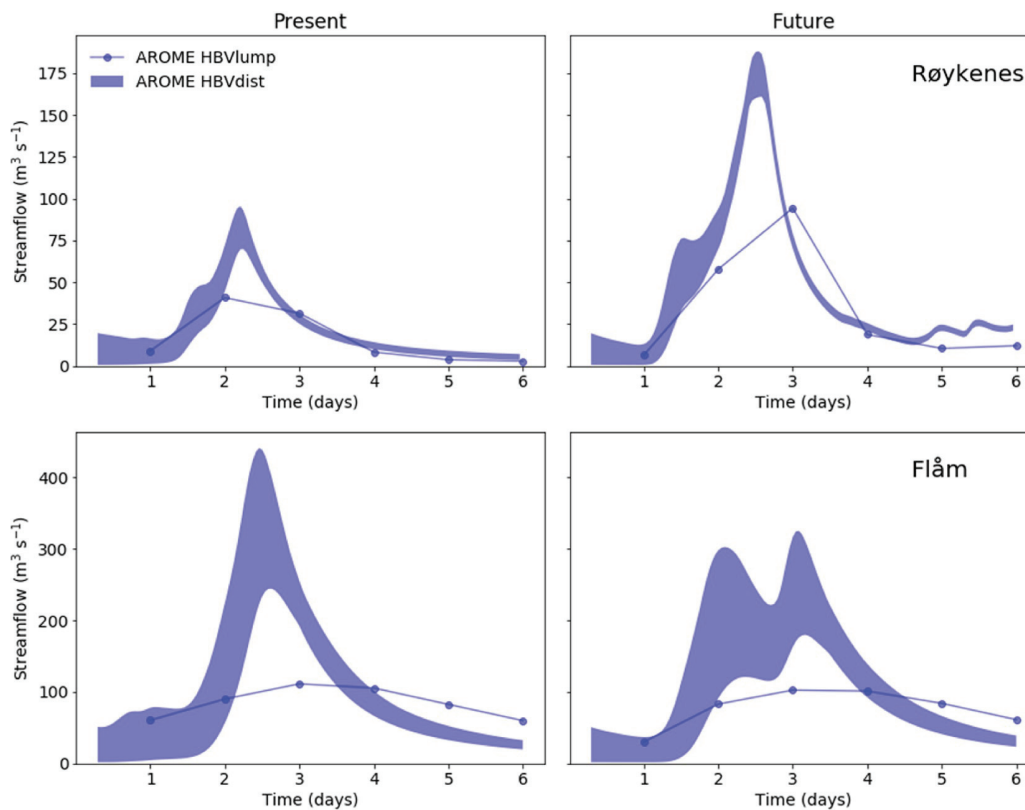


**Fig. 6.** The simulated streamflow for Røykenes (top) and Flåm (bottom), in the present (left) and future (right), using HBVlump and HBVdist. The HBVlump streamflow simulations are the same as in Fig. 5, but only the AROME present and future ensemble member resulting in the highest streamflow peak is shown. Similarly, for HBVdist, only the AROME present and future ensemble member that results in the highest streamflow peak are selected, and the uncertainty range represents hydrologic uncertainty (section 2.4).

for both the present and future events in both catchments. The simulations are 3–4 (Flåm) and 1 to 2 (Røykenes) times higher for HBVdist than HBVlump. The streamflow simulations show clearly the influence of hourly input in HBVdist showing more temporal variability than the HBVlump with daily input. For example, for Røykenes, the streamflow between the 2nd and 3rd day of the event is similar for HBVdist and HBVlump, whereas the peaks, occurring around 2.5 days, are much higher for HBVdist. The hourly peak compared to the daily peak for

Røykenes is within the expected values when looking up the factor used to estimate the flood culmination value based on the daily flood value. For Flåm, however, the difference between hourly and daily peak is too large and therefore it is difficult to compare the results directly. The large discrepancy between the HBVdist and HBVlump could be caused by the fact that the models are calibrated on different versions of the seNorge dataset. The hourly data are available for a shorter period and less quality checked than the daily data and different interpolation

methods are applied that affect the precipitation (v2.0 has less precipitation amounts than v1.1). Both can have a large effect on the tuning of hydrological model parameters (connected to precipitation) and thereby the estimation of floods.

In the previous section it was shown that the local, catchment-scale impact can be less in future events compared to present events depending on the catchment. This effect was predominantly found for the AROME events, because the high-resolution weather forecast model shows more spatial variability than the higher resolution EC-Earth model. The fact that some catchments might and others might not be hit by a precipitation event is consistent with streamflow observations of the 2014 event, which show that Flåm was hit but Røykenes, among others, was not (SI Fig. 10). As mentioned before, Fig. 6 shows only the AROME event resulting in the highest streamflow rather than the average of the 10 events, but leads to the same conclusions drawn from Figs. 3 and 4. That is, peak streamflows in Røykenes are higher in the future compared to present, whereas in the Flåm catchment, the future event is weaker than in present. The future event in Røykenes shows an almost two-fold increase in streamflow at peak-time in addition to extending over a longer period of time (Fig. 6). The peak streamflow of ca. 180 m$^3$/s$^1$ of this event is ca. 20% higher compared to the largest flood event ever measured in this catchment, a flood in 1953 with a streamflow peak of ca. 140 m$^3$/s$^1$ (Væringstad, 2014).

## 4. Conclusions

After consulting a range of users and stakeholders, the autumn floods of September 2005 and October 2014, both caused by an atmospheric river hitting the West Coast of Norway, were chosen as a typical event for this physical climate storyline study. In particular, we investigate the impact of having higher spatial (from 25 km in EC-Earth to 2.5 km in AROME) and temporal resolution (daily vs. hourly) for the simulation of peak flows in two catchments. In addition, we are using a modelling chain that mimics as closely as possible the operational modelling chain for flood warnings in Norway.

We first show that the present event is realistically simulated by both EC-Earth and AROME when looking at a larger region in the West Coast of Norway. EC-Earth's events are also slightly wetter than in AROME, but this is likely an artefact of the different resolutions. AROME shows stronger local peaks in precipitation but also some locations appear to receive almost no precipitation during the event, which is plausible in such a mountainous area. EC-Earth, in contrast, shows more homogenous precipitation amounts for the larger grid cells and misses the fact that extreme precipitation actually falls in a more localised manner. The high resolution of the EC-Earth simulations (compared to e.g. CMIP5 climate models) allow for more realistic atmospheric circulation patterns (e.g. van Haren et al., 2015), but for applications studied here, the resolution is insufficient. Downscaling with AROME is necessary to obtain more realistic precipitation distributions. In both EC-Earth and AROME however, there are differences between the 10 ensemble members, which can likely be traced back to where the atmospheric river is located, and how intense it is in the perturbed EC-Earth reruns. The difference between present and future events follow overall the expectation that extreme precipitation will mostly increase in this region in Autumn (Hanssen-Bauer et al., 2017), while mean precipitation is expected to decrease in the south and increase on the northern West Coast. Whan et al. (2019)'s results suggest that this increase in extreme precipitation is caused by more intense and frequent atmospheric rivers hitting the region in the future in October–March, which should be taken into account in future adaptation planning in the region.

We then investigate the local impact on two catchments at the West Coast of Norway: Flåm and Røykenes. Whereas EC-Earth appears overall wetter when considering the larger region of the West Coast of Norway, we find that AROME produces more precipitation (in terms of intensity and quantity) for the two studied catchments. Translating this to physical impacts, the AROME simulations result in higher peak streamflows

compared to EC-Earth. The difference between present and future events shows a future increase in streamflow for Røykenes, and a decrease for Flåm. This result is consistent for the two hydrological model versions analysed, with the higher resolution model simulating even larger peak streamflows. Previous studies also showed that for the West Coast, autumn peak streamflow is expected to become stronger with precipitation increases in a warmer climate (Beldring, 2008; Hanssen-Bauer et al., 2017; Lawrence and Hisdal, 2011; Sorteberg et al., 2018; Vormoor et al., 2015). The unexpected result that the future events in Flåm have less physical impact (i.e. less streamflow) is likely due to the fact that the future event did not hit Flåm. Even though the streamflow peaks are on average higher in a future climate, not every future event will hit every catchment. We attempted to take into account the uncertainty regarding which catchments are hit by perturbing each selected event ten times (see section 2). Our selected cases cannot span the entire range of possibilities where an AR reaches the coast of Norway or its passage over land, but we can go into much more detail for a specific case. Physical climate storylines should be viewed as "what if" scenarios that can provide insights, for instance for emergency preparedness, even if the probability of that event is unknown or very unlikely, and to study possible impacts associated with such an event. Physical climate storylines can be used to reveal complexity and uncertainties in the system that cannot be revealed in a coarser, more probabilistic focused approach.

Overall, Hegdahl et al. (2020), found that the future events cause larger floods coinciding in many more catchments compared to the present events, which underlines that the severity of an event may not necessarily be captured in a single catchment analysis. We find that hourly timescales (instead of daily) represent the fast flood-generating processes in these catchments better and therefore provide a more realistic temporal evolution of the flood event, capturing higher peak flood intensities. We conclude that especially in a mountainous region such as the West Coast of Norway, higher spatial and temporal resolutions in both the climate and hydrological models lead to more realistic estimations of the potential peak floods, which is important information for users, such as NVE. Another valuable aspect of physical climate storylines is the familiarity of the operational setup and data format for the users. NVE and Statkraft could both easily use the data produced.

When considering individual events as is the case in a storyline approach, it is obviously impossible to make any statements about their probability of occurrence. It is however possible to put them in context of expected future climate change from other GCM ensembles such as CMIP5 or from the literature, as we did here. Also, it is possible to study the dynamics of the event and possible compounded effects in detail, which are often obscured in probabilistic projections. At the same time, it is also key to communicate that each event is unique, and no general "projection" can be obtained from a storyline approach. In addition, communicating different outcomes for different catchments to the users, and in our case, the counterintuitive results that the future event in Flåm is weaker compared to the present one, is not an easy task. The community needs to be aware that physical climate storylines studies might not always lead to the expected results, but having the users involved from the design of the storyline until the interpretation of results can support mutual understanding of the process and challenges involved. Further studies are needed to better understand the added value of a physical climate storyline approach for users, such as municipalities challenged with making decision for adaptation measures for future climate change (Bremer et al., 2019; Cortes Arevalo et al., 2018; de Bruijn et al., 2016).

Physical modelling supporting a storyline approach is a relatively new field of research in the physical climate change community and a lot of research is ongoing and needed (Sillmann et al., 2019). We briefly summarise here a few key aspects discovered in this study that require further research:

1) As in probabilistic event attribution, the definition of the event is highly important (Otto et al., 2015). For example, in hindsight it is clear that the selection criteria for the event in the EC-Earth PRESENT and FUTURE simulations (i.e. 24 h accumulated precipitation) could have been optimized by considering longer event duration (e. g. 3 days) which will likely affect the results obtained further down in the modelling chain. Sensitivity analysis could be done given sufficient computational resources that were not available for this study. However, our study was able to show that high spatial resolution throughout the model chain (including the global climate model) is needed for realistically simulating the magnitude of AR events and associated floods.

2) As Felder et al. (2018) also discuss, there is no guarantee that the chosen event in the GCM will correspond to an extreme event once downscaled. The catchment of Flåm is clearly not much affected in the future event, leading to the fact that the present event appears stronger, which is not trivial to communicate with the users.

3) We performed 10 re-runs with the driving GCM before doing the downscaling, using one way of perturbing the initial conditions, but more research should be done on this to know how important internal variability is. We chose 10 ensemble members as a trade-off between computing capacity and representation of internal variability, but there is no physical or statistical reason beyond this.

4) In our region of interest, the West Coast of Norway, temperature is important for the partitioning between snow and rain, and snowmelt when there is snow in the catchment. The flood development during cold seasons are therefore also depending on temperature. If precipitation falls and is stored as snow, it will not contribute to streamflow. In this study we focussed on October when temperatures are usually still above the freezing point. The impact of a future change in temperature is not only important to the amount of precipitation reaching western Norway, but also for the phase of the precipitation (Poschlod et al., 2020). A winter AR which today deposits snow at higher elevations can in a warmer climate cause floods (Whan et al., 2019).

## Declaration of competing interest

The authors declare no conflict of interests.

## CRediT authorship contribution statement

**Nathalie Schaller:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Visualization. **Jana Sillmann:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Malte Müller:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Funding acquisition. **Reindert Haarsma:** Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing - original draft. **Wilco Hazeleger:** Conceptualization, Methodology, Writing - review & editing, Project administration, Funding acquisition. **Trine Jahr Hegdahl:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Visualization. **Timo Kelder:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Visualization. **Gijs van den Oord:** Methodology, Validation. **Albrecht Weerts:** Methodology, Writing - review & editing, Supervision. **Kirien Whan:** Methodology, Validation, Formal analysis, Writing - review & editing, Visualization.

## Acknowledgement

We thank Noelia Otero for her help with calculating the weather types.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wace.2020.100259.

## References

Azad, R., Sorteberg, A., 2017. Extreme daily precipitation in coastal western Norway and the link to atmospheric rivers. J. Geophys. Res. Atmos. 122, 2080–2095. https://doi.org/10.1002/2016JD025615.

Beldring, S., 2008. Distributed element water balance model system. NVE Report 2008, 42pp, 4. http://publikasjoner.nve.no/report/2008/report2008_04.pdf.

Beldring, S., Engen-Skaugen, T., Førland, E.J., Roald, L.A., 2008. Climate change impacts on hydrological processes in Norway based on two methods for transferring regional climate model results to meteorological station sites. Tellus Dyn. Meteorol. Oceanogr. 60 (3), 439–450.

Benedict, I., Ødemark, K., Nipen, T., Moore, R., 2019a. Large-scale flow patterns associated with extreme precipitation and atmospheric rivers over Norway. Mon. Weather Rev. https://doi.org/10.1175/MWR-D-18-0362.1.

Benedict, I., van Heerwaarden, C.C., Weerts, A.H., Hazeleger, W., 2019b. The benefits of spatial resolution increase in global simulations of the hydrological cycle evaluated for the Rhine and Mississippi basins. Hydrol. Earth Syst. Sci. 23, 1779–1800. https://doi.org/10.5194/hess-23-1779-2019.

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., Santos Muñoz, D., Subias, A., Tijm, S., Toll, V., Yang, X., Køltzow, M.Ø., 2017. The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. Mon. Weather Rev. 145, 1919–1935. https://doi.org/10.1175/MWR-D-16-0417.1.

Bergström, S., 1976. Development and application of a conceptual runoff model for Scandinavian catchments. In: SMHI RHO, vol. 7. Norrköping.

Bremer, S., Wardekker, A., Dessai, S., Sobolowski, S., Slaattelid, R., van der Sluijs, J., 2019. Toward a multi-faceted conception of co-production of climate services. Climate Services 13, 42–50. https://doi.org/10.1016/j.cliser.2019.01.00.

Cortes Arevalo, V.J., Verbrugge, L.N.H., Haan, R.-J. d., Baart, F., van der Voort, M.C., Hulscher, S.J.M.H., 2018. Users' perspectives about the potential usefulness of online storylines to communicate river research to a multi-disciplinary audience. Environ. Commun. 13 (7), 909–925. https://doi.org/10.1080/17524032.2018.1504098.

de Bruijn, K.M., Lips, N., Gersonius, B., Middelkoop, H., 2016. The storyline approach: a new way to analyse and improve flood event management. Nat. Hazards 81 (1), 99–121. https://doi.org/10.1007/s11069-015-2074-2.

Dee, D.P., et al., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597.

Dyrrdal, A.V., Skaugen, T., Stordal, F., Førland, E.J., 2016. Estimating extreme areal precipitation in Norway from a gridded dataset. Hydrol. Sci. J. 61, 483–494. https://doi.org/10.1080/02626667.2014.947289.

Felder, G., Gómez-Navarro, J.J., Zischg, A.P., Raible, C.C., Röthlisberger, V., Bozhinova, D., Martius, O., Weingartner, R., 2018. From global circulation to local flood loss: coupling models across the scales. Sci. Total Environ. 635, 1225–1239. https://doi.org/10.1016/j.scitotenv.2018.04.170.

Haarsma, R.J., Hazeleger, W., Severijns, C., de Vries, H., Sterl, A., Bintanja, R., van Oldenborgh, G.J., van den Brink, H.W., 2013. More hurricanes to hit western Europe due to global warming. GRL 40 (9), 1783–1788.

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., Ådlandsvik, B., 2017. Climate in Norway 2100 - a Knowledge Base for Climate Adaption. NCCS Report 1/2017. https://www.miljodirektoratet.no/globalassets/publikasjoner/M741/M741.pdf.

van Haren, R., Haarsma, R.J., van Oldenborgh, G.J., Hazeleger, W., 2015. Resolution dependence of European precipitation in a state-of-the-art atmospheric general circulation model. J. Clim. 28.

Hazeleger, W., et al., 2010. EC-earth A seamless earth-system prediction approach in action. BAMS. https://doi.org/10.1175/2010BAMS2877.1.

Hazeleger, W., van den Hurk, B.J.J.M., Min, E., van Oldenborgh, G.J., Petersen, A.C., Stainforth, D.A., Vasileiadou, E., Smith, L.A., 2015. Tales of future weather. Nat. Clim. Change 5. https://doi.org/10.1038/NCLIMATE2450.

Hegdahl, Trine, Engeland, Kolbjørn, Müller, Malte, Sillmann, Jana, 2020. An event-based approach to explore selected present and future Atmospheric River induced floods in western Norway. Journal of Hydrometeorology. https://doi.org/10.1175/JHM-D-19-0071.1.

Hodnebrog, Ø., Marelle, L., Alterskjær, K., Wood, R.R., Ludwig, R., Fischer, E.M., Richardson, T.B., Forster, P.M., Sillmann, J., Myhre, G., 2019. Future changes in mean and extreme summer precipitation over Europe. ERL 14, 124050.

Houska, T., et al., 2015. Spotting model parameters using a ready-made python package. In: Langsholt, E., Roald, L.A., Holmqvist, E., Fleig, A. (Eds.), 2015. Flommen På Vestlandet Oktober 2014. NVE Rapport Nr 112015.

van den Hurk, B.J.J.M., Viterbo, P., Beljaars, A.C.M., Betts, A.K., 2000. Offline validation of the ERA40 surface scheme. ECMWF TechMemo 295. http://www.ecmwf.int/publications/library/ecpublications/pdf/tm/001-300/tm295.pdf.

Iden, K, Isaksen, K, Kristiansen, S, Szewczyk-Bartnicka, H, 2005. Weather in Norway-Climatological monthly overview for September 2005. ), met.no.info (09/2005), Meteorological Institute, Oslo, Norway. ISSN 1503-8017.

Jacob, D., et al., 2014. EURO-CORDEX: new high-resolution climate change projections for European impact research. Reg. Environ. Change 14, 563–578.

Lawrence, D., Hisdal, H., 2011. Hydrological projections for floods in Norway under a future climate. Nor. Water Resour. Energy Dir. Rep. 2011 (5). ISBN: 978-82-410-0753-8.

Lussana, C., Saloranta, T., Skaugen, T., Magnusson, J., Tveito, O.E., Andersen, J., 2018. seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the present day. Earth Syst. Sci. Data 10, 235–249 https://doi.org/10.5194/essd-10-235-2018.

Mohr, M., 2008. New Routines for Gridding of Temperature and Precipitation Observations for "seNorge. No". MET report No 08/2008, available at. ftp://ftp.met.no/projects/klimagrid/doc/NewRoutinesforGriddingofTemperature.pdf.

Mohr, M., 2009. Comparison of Versions 1.1 and 1.0 of Gridded Temperature and Precipitation Data for Norway. MET report No 19/2009, available at. ftp://157.2 49.32.246/projects/klimagrid/doc/Comparison%20of%20versions%201.1%20and %201.0%20of%20gridded%20climate%20data.pdf.

Müller, M., Homleid, M., Ivarsson, K.-I., Køltzow, M.A.Ø., Lindskog, M., Midtbø, K.H., Andrae, U., Aspelien, T., Berggren, L., Bjørge, D., Dahlgren, P., Kristiansen, J., Randriamampianina, R., Ridal, M., Vignes, O., 2017. AROME-MetCoOp: a Nordic convective-scale operational weather prediction model. Weather Forecast. 32, 609–627. https://doi.org/10.1175/WAF-D-16-0099.1.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part i—a discussion of principles. J. Hydrol. 10 (3), 282–290.

Otero, Noelia, Sillmann, Jana, Butler, Tim, 2018. Assessment of an extended version of the Jenkinson-Collison classification on CMIP5 models over Europe. Climate Dynamics 50. https://doi.org/10.1007/s00382-017-3705-y.

Otto, F.E.L., Boyd, E., Jones, R.G., Cornforth, R.J., James, R., Parker, H.R., Allen, M.R., 2015. Attribution of extreme weather events in Africa: a preliminary exploration of the science and policy implications. Climatic Change. https://doi.org/10.1007/s10584-015-1432-0.

Porter, J.J., Dessai, S., 2017. Mini-me: why do climate scientists' misunderstand users and their needs? Environ. Sci. Pol. 77, 9–14.

Poschlod, B., Zscheischler, J., Sillmann, J., Wood, R.R., Ludwig, R., 2020. Climate change effects on hydrometeorological compound events over southern Norway. Weather. Clim. Extr. 28, 100254.

Prein, A.F., et al., 2015. A review on regional convection-permitting climate modeling: demonstrations, prospects, and challenges. Rev. Geophys. 53, 323–361.

Rogelj, J., Meinshausen, M., Knutti, R., 2012. Global warming under old and new scenarios using IPCC climate sensitivity range estimates. Nat. Clim. Change 2, 248–253. https://doi.org/10.1038/nclimate1385.

Sælthun, N.R., 1996. The Nordic HBV model. Nor. Water Resour. Energy Dir. 7, 1–26.

Sandvik, M.I., Sorteberg, A., Rasmussen, R., 2018. Sensitivity of historical orographically enhanced extreme precipitation events to idealized temperature perturbations. Clim. Dynam. 50, 143–157. https://doi.org/10.1007/s00382-017-3593-1.

Schaller, N., Kay, A.L., Lamb, R., Massey, N.R., van Oldenborgh, G.J., Otto, F.E.L., Sparrow, S.N., Vautard, R., Yiou, P., Ashpole, I., Bowery, A., Crooks, S.M., Haustein, K., Huntingford, C., Ingram, W.J., Jones, R.G., Legg, T., Miller, J., Skeggs, J., Wallom, D., Weisheimer, A., Wilson, S., Stott, P.A., Allen, M.R., 2016. Human influence on climate in the 2014 Southern England winter floods and their impacts. Nat. Clim. Change 6 (6).

Schellekens, J., van Verseveld, W., de Boer-Euser, T., Winsemius, H., Thiange, C., Bouaziz, L., Tollenaar, D., de Vries, S., Weerts, A.H., 2017. Openstreams wflow documentation. available at: https://wflow.readthedocs.io/en/stable/. .

Sharma, A., Wasko, C., Lettenmaier, D.P., 2018. If precipitation extremes are increasing, why aren't floods? Water Resour. Res. 54, 8545–8551.

Shepherd, T.G., 2019. Storyline approach to the construction of regional climate change information. Proc. R. Soc. A 475, 20190013. https://doi.org/10.1098/rspa.2019.0013.

Shepherd, T.G., Boyd, E., Calel, R.A., Chapman, S.C., Dessai, D., Dima-West, I.M., Fowler, H.J., James, R., Maraun, D., Martius, O., Senior, C.A., Sobel, A.H., Stainforth, D.A., Tett, S.F.B., Trenberth, K.E., van den Hurk, B.J.J.M., Watkins, N.W., Wilby, R.L., Zenghelis, D.A., 2018. Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. Climatic Change 151, 555–571.

Sillmann, J., Kharin, V.V., Zwiers, F.W., Zhang, X., Bronaugh, D., 2013. Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. J. Geophys. Res. Atmos. 118, 2473–2493. https://doi.org/10.1002/jgrd.50188.

Sillmann, J., Shepherd, T., van den Hurk, B., Hazeleger, W., Martius-Rompainen, O., Zscheischler, J., 2019. Physical modeling supporting a storyline approach. CICERO Policy Note 2019:01, available at: https://www.cicero.oslo.no/en/publications-an d-events_twex/workshop-on-physical-modeling-supporting-a-storyline-approach.

Sodemann, H., Stohl, A., 2013. Moisture origin and meridional transport in atmospheric rivers and their association with multiple cyclones. Mon. Weather Rev. 141, 2850–2868. https://doi.org/10.1175/MWR-D-12-00256.1.

Sorteberg, A., Lawrence, D., Dyrrdal, A.V., Mayer, S., Engeland, K. (Eds.), 2018. Climate Changes in Short Duration Extreme Precipitation and Rapid Onset Flooding – Implications for Design Values. Norwegian Center for Climate Services. NCCS Report 2018:1, available at: https://cms.met.no/site/2/klimaservicesenteret/rapporte r-og-publikasjoner/_attachment/13537?_ts=163df95ff7b.

Sterl, A., Severijns, C., Dijkstra, H., Hazeleger, W., van Oldenborgh, G.J., van den Broeke, M., Burgers, G., van den Hurk, B., van Leeuwen, P.J., van Velthoven, P., 2008. When can we expect extremely high surface temperatures? Geophys. Res. Lett. 35, L14703.

Stohl, A., Forster, C., Sodemann, H., 2008. Remote sources of water vapor forming precipitation on the Norwegian West Coast at 60ºN—a tale of hurricanes and an atmospheric river. J. Geophys. Res. 113, D05102. https://doi.org/10.1029/2007JD009006.

Swedish Meteorological and Hydrological Institute, 2014. IHMS - Integrated Hydrological Modelling System. Manual version 6.4.

Taylor, K.E., Stouffer, R.J., Meehl, G.A., 2012. An overview of CMIP5 and the experiment design. Bull. Am. Meteorol. Soc. 93, 485–498.

Væringstad, T., 2014. Flomberegning for Nesttunvassdraget (056.3Z), Rapport Nr. 16 – 2014. NVE Report, available at: http://publikasjoner.nve.no/rapport/2014/rappor t2014_16.pdf.

Vormoor, K., Lawrence, D., Heistermann, M., Bronstert, A., 2015. 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, 913–931. https://doi.org/10.5194/hess-19-913-2015.

van Vuuren, D.P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G.C., Kram, T., Krey, V., Lamarque, J.F., Masui, T., Meinshausen, M., Nakicenovi, N., Smith, S.J., Rose, S.K., 2011. The representative concentration pathways: an overview. Climatic Change 109, 5–31.

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., Heynert, K., 2013. The Delft-FEWS flow forecasting system. Environ. Model. Software 40, 65–77. https://doi.org/10.1016/j.envsoft.2012.07.010.

Whan, K., Sillmann, J., Schaller, N., Haarsma, R., 2019. Future changes in atmospheric rivers and extreme precipitation in Norway. Clim. Dynam. 54, 2071–2084.