https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

The ability of LSTM to model snowmelt versus rainfall generated floods

Sigrid Jørgensen Bakke[1], Danielle Marie Barna[1], Kolbjørn Engeland[1], Sjur Anders Kolberg[1], and Sunniva Nordeide[1]

[1]Department of Hydrology, Norwegian Water Resources and Energy Directorate, Oslo, Norway

Correspondence: Sigrid Jørgensen Bakke (sijb@nve.no)

Abstract. One of the most important skills of hydrological models is to simulate timing and magnitude of flood events. Long Short-Term Memory (LSTM) networks are currently among the most successful models for streamflow and flood prediction over large regions. In snow-influenced catchments, which typically comprise a minority in large-scale studies, floods are generated by two distinctly different processes, snowmelt and rainfall. The applicability of hydrological models in such regions is therefore dependent on their ability to represent both types of floods. Nevertheless, flood evaluations of LSTM taking different flood-generating processes into account are currently lacking. This study fills this gap by evaluating the ability of LSTM to model flood peak characteristics separately for snowmelt and rainfall generated floods. The trained LSTM model successfully simulated streamflow time series across the 103 evaluated catchments, with average NSE of 0.85 and average KGE of 0.87 over the unseen evaluation period. LSTM exhibited better performance in the majority of the catchments in terms of flood peak timing and magnitude for both rainfall and snowfall generated floods when compared to the operational hydrological model in the region (HBV) used as a benchmark. Both models had a 24 pp higher percentage of correctly simulated peak days for rainfall generated floods as compared to snowmelt generated floods. LSTM outperformed HBV for a larger proportion of the catchments in terms of peak timing of rainfall generated events (83 %) as compared to snowmelt generated events (64 %). On the other hand, a larger proportion of the catchments were improved by LSTM for snowmelt generated events as compared to rainfall generated events when considering peak magnitudes. The largest improvements in peak magnitudes were found for rainfall generated events, in particular for catchments where HBV exhibited high (>40 %) absolute errors. Overall, our findings bring confidence that LSTM can improve hydrological services in regions subject to both snowmelt and rainfall generated floods.

1 Introduction

Reliable flood simulations are important for a range of national to local tasks, including area planning, flood and landslide forecasting, hydropower management, and climate change impact assessments. Process-based hydrological models that most commonly perform this task, have recently been challenged by deep learning Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997; Gers et al., 2000). Variants of LSTM networks have proven their success as rainfall-runoff models across regions, often outperforming benchmark process-based models in terms of streamflow prediction (Kratzert et al.,

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

2018), riverine flood prediction (Frame et al., 2022) and forecasting (Nearing et al., 2024). Unlike most process-based models that perform the best when they are locally calibrated, LSTM, being a purely data-driven model, excels when trained on big datasets of multiple catchments (Kratzert et al., 2024, 2019b). Thus, LSTM has potential to both improve and simplify hydrological services on local to global scales.

In mountainous and northern regions, it is particularly important that an applied hydrological model is able to capture effects of seasonal snow accumulation and melt on streamflow. Without physical processes or constraints explicitly implemented in the model code, a purely data-driven model needs to learn such effects solely from the training data. In LSTM, long-term dependencies between input and output time series are handled by memory cells that can represent depletion, increase and outflow of reservoirs and storages in the case of hydrological modelling (Kratzert et al., 2019a, 2018). Studies have indeed shown how LSTM cell state vectors produce temporal dynamics matching our physical understanding of snow accumulation and melt (Kratzert et al., 2018), and with good (>0.8) correlations with snow depth reanalysis products (Lees et al., 2022). Correspondingly, LSTMs have demonstrated good performance in snow-influenced regions in terms of simulating streamflow in general (Roksvåg et al., 2026; Jiang et al., 2022), and flood peaks in particular (Martel et al., 2025; Hagen et al., 2023; Anderson and Radić, 2022).

As snowmelt and rainfall generated floods are substantially different in their underlying physics and temporal dependencies of past weather, the performance of a hydrological model may differ for the two types of floods. Identifying and understanding such difference is important both for improving models and quantification of uncertainty in model simulations and forecasts. Nevertheless, to our knowledge, existing evaluations of LSTM's ability to simulate floods are performed over all identified flood events, without a separation into events generated by different processes. Evaluation results over all flood peaks mainly reflect the model's ability to simulate the flood type possessing the majority of the events, whereas the ability to simulate other flood types of importance is left unknown (Martel et al., 2025). This is a particular issue for regions with seasonal snow cover where floods are primarily driven by two fundamentally different processes, snowmelt and rainfall, and an applied hydrological model should be able to simulate both flood generating processes.

This study meets the abovementioned gap by investigating the ability of LSTM to simulate floods of different driving mechanisms. The main objective is to evaluate LSTM's potential for operational use in snow-influenced regions, focusing on prediction of floods dominantly generated by snowmelt and rainfall separately. The study area is the mainland of Norway, a country spanning 58-71 degrees north that has a large variability in topography, annual precipitation amounts and seasonal snow cover. The applicability of LSTM for a given country or region depends on the performance of LSTM versus the regions' state-of-the-art operational hydrological model in terms of hydrological features that are important in that region. Accordingly, we use existing simulations from a national operational hydrological model as a benchmark. We meet our aim by answering the following research questions:

1. Is LSTM able to simulate streamflow series with similar or exceeding overall performance as compared to the operational model used in the region?

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

2. How well does LSTM simulate peak timing and magnitude of all collected floods generated by i) snowmelt, ii) rainfall and iii) a mix of the two?

3. How well does LSTM simulate per-catchment peak timing and magnitude of i) snowmelt generated floods and ii) rainfall generated floods?

The paper is structured as follows: In Sect. 2 the data underlying the analyses are described, followed by methods explaining the LSTM model set-up and training, flood event detection and classification, and the applied model evaluation metrics. Section 3 presents the results, which are further discussed in Sect. 4 before conclusions are drawn in Sect. 5.

2 Data and methods

2.1 Discharge data and catchment attributes used for LSTM

Daily discharge data used to train the LSTM network stemmed from 200 streamflow gauging stations of the Norwegian Water Resources and Energy Directorate (NVE) hydrometric observation network in Norway (Fig. 1). Of those stations, model evaluation was performed over the 103 stations with available discharge simulations from the operational model used as a benchmark (ref. Sect. 2.3). Each day in the observed daily discharge series is divided at midnight Central European Time (CET).

The gauging stations were selected from an existing dataset that has been thoroughly quality controlled for flood analyses (530 stations; Engeland et al., 2016). Requirements during that selection included: The stations should not have problems with ice jams or supercritical flow during floods, they should have satisfactory rating curves and measurement conditions during floods, and they should not have homogeneity breaks for streamflow and floods that stemmed from changes in the measurements or quality control. Of the 530 stations, we selected the active stations unaffected by regulations. Further, only stations with no more than one year of maximum 60 missing days in the calibration period (i.e. 2004–2018) were selected as a compromise between the number of stations and period coverage. The original unit of $m^3\ s^{-1}$ was changed to mm day^{-1} by normalising the streamflow values by the catchment areas. Catchment areas represented in the dataset range 0.58 km^2 up to 14 000 km^2. However, most areas are less than 300 km^2, which was expected due to the requirement of near-natural flow in a country heavily affected by hydropower.

Table 1 lists the catchment attributes selected for the training of the LSTM model. The 21 catchment attributes were selected to represent the variability of geometric, physiographic and climatological characteristics important for streamflow, hydrological regimes and floods. There is a large degree of overlap between our selection and attributes that have previously been selected for regional flood frequency models in Norway (Engeland et al., 2020; Barna et al., 2023). Most of the catchments are influenced by seasonal snow. Of all the 200 catchments, 95 % have average winter (DJF) temperatures below 0 °C (range -13 to 3 °C), whereas average summer (JJA) temperatures are all positive, ranging 6 to 16 °C. Several of the catchments are also partly covered by glaciers.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

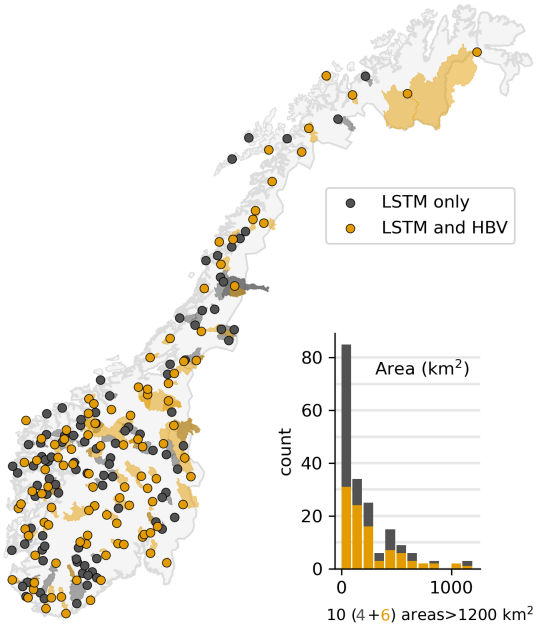

Figure 1. Locations and catchments of the streamflow gauging stations, and histogram of catchment areas. All 200 catchments (both yellow and dark grey) were used to train the LSTM model, whereas yellow represents the 103 catchments where both LSTM and HBV have simulations (used for model evaluation).

2.2 Gridded hydrometeorological data

Daily precipitation sum (mm) and mean daily temperature (°C) from the meteorological dataset SeNorge_2018 (Lussana et al., 2019) was used to produce the dynamical forcing for LSTM (Sect. 2.5). Flood event classification (Sect. 2.6) was based on the same precipitation data, as well as daily snowmelt (mm) estimates simulated by the SeNorge snow model v.1.1.1 (Saloranta, 2016) that uses SeNorge_2018 gridded daily precipitation sum and mean temperature as input. Both SeNorge_2018 and the snowmelt dataset are available from 1957 to present at a daily time step (07:00–07:00 CET). They are gridded products with a spatial resolution of 1x1 km^2, covering the whole of Norway as well as neighbouring regions. The SeNorge_2018 dataset is generated by statistical interpolation of in situ precipitation and temperature observations. The SeNorge snow model used for classifying flood events simulates snowmelt by a degree-day approach augmented with synthetic incoming daily solar irradiance defined as a function of latitude and time of the year.

From the daily gridded variables precipitation sum and mean temperature, the following daily time series were prepared at the catchment level for LSTM modelling:

- prec (mm day^{-1}): Spatially averaged daily precipitation sum

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

Table 1. Catchment attribute names, explanations, units and references to data.

Name	Explanation	Unit	Reference to data
area	Catchment area	km^2	GeoNorge (2026a)
areaperperimeter_km	Area divided by circumference	km	Derived from GeoNorge (2026a)
height_hypso_10	10 % percentile of hypsographic curve	masl	NVE's database (available at NVE, 2026b)
height_hypso_90	90 % percentile of hypsographic curve	masl	NVE's database (available at NVE, 2026b)
length_km_river	Length of main river	km	GeoNorge (2026b)
slope_mean	Catchment mean slope	degrees	Kartverket (2026)
gradient_1085	Gradient of main river excluding the 10 % lowest and 15 % highest reaches	m km^{-1}	NVE's database, based on GeoNorge (2026b)
drainage_density	Drainage density (total river length divided by catchment area)	km^{-1}	NVE's database, based on GeoNorge (2026b)
infp	Infiltration potential (1–5)	-	GeoNorge (2026c)
perc_forest	Percentage of catchment covered by forest	%	NVE's database (available at NVE, 2026b)
perc_eff_lake	Effective lake percentage, defined as $100 * \sum_l^L (A_l \ a_l)/A$, where A is catchment area, and a_l and A_l are surface area and drainage area of lake $l = 1, ..., L$, respectively	%	NVE's database
perc_glacier1985	Percentage of catchment covered by glaciers in 1985	%	Winsvold et al. (2014)
perc_glacier2019	Percentage of catchment covered by glaciers in 2019	%	Andreassen et al. (2022)
t_djf	Mean winter (DJF) temperature 1991–2020	°C	Tveito (2021)
t_mam	Mean spring (MAM) temperature 1991–2020	°C	Tveito (2021)
t_jja	Mean summer (JJA) temperature 1991–2020	°C	Tveito (2021)
t_son	Mean autumn (SON) temperature 1991–2020	°C	Tveito (2021)
p_djf	Mean winter (DJF) precipitation sum 1991–2020	mm $winter^{-1}$	Tveito (2021)
p_mam	Mean spring (MAM) precipitation sum 1991–2020	mm $spring^{-1}$	Tveito (2021)
p_jja	Mean summer (JJA) precipitation sum 1991–2020	mm $summer^{-1}$	Tveito (2021)
p_son	Mean autumn (SON) precipitation sum 1991–2020	mm $autumn^{-1}$	Tveito (2021)

- temp_ave (°C): Spatially averaged daily mean temperature
- temp_min (°C): The minimum value of all catchment grid cells' daily mean temperature
- temp_max (°C): The maximum value of all catchment grid cells' daily mean temperature

Further, based on daily gridded precipitation sum and snowmelt, the following time series were prepared at the catchment level to identify flood generating processes:

- snowmelt (mm day^{-1}): Spatially averaged daily snowmelt
- rainfall (mm day^{-1}): Spatially averaged daily rainfall sum. For each grid cell, the daily rainfall sum was defined as the daily precipitation sum if the daily mean temperature was equal or larger than 0.5 °C, and 0 mm day^{-1} otherwise.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

2.3 Benchmark model

The Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström, 1976) is a widely used precipitation-runoff model, in particularly in the Nordic countries (Addor and Melsen, 2019; Seibert and Bergström, 2022). In Norway, a version of the HBV model with a daily temporal resolution is used operationally by the Norwegian Water Resources and Energy Directorate (NVE), for national flood warning services, quality control of streamflow observations, and energy prognosis among others (Lawrence et al., 2009; Ruan and Langsholt, 2017). For this reason, we chose existing simulations from the operational HBV model for benchmarking our results. HBV simulations cover the period 1957–2023 and exist for 103 of the 200 gauging stations used for LSTM. Thus, all model evaluations were done for those 103 stations (Fig. 1).

For descriptions of the version and set-up of the daily HBV model used operationally in Norway, we refer to Lawrence et al. (2009) and Ruan and Langsholt (2017), except that the model has been re-calibrated for a more recent period (2004–2018) using a more recent meteorological dataset (SeNorge_2018) after the two reports were published. In short, the HBV model underlying the applied simulations is a semi-distributed bucket-type model, that divides the catchment into ten elevation zones to account for elevation gradients in temperature and precipitation (Sælthun, 1996; Killingtveit and Saelthun, 1995). The model uses elevation zone averaged daily precipitation sums and mean temperatures as input. Fluxes between and processes within four storage components, i.e. snow, soil moisture, an upper runoff zone, and a lower runoff zone, are represented by simplified process-based expressions. Catchment-by-catchment calibration of 12 model parameters were conducted using Model-Independent Parameter Estimation and Uncertainty Analysis (PEST; Doherty, 2004) by minimizing the sum of squared errors plus squared total volume error. The operational HBV model has been locally calibrated for 103 of the 200 catchments used to train LSTM (Fig. 1).

Daily discharge and meteorological observations in Norway have historically been recorded at two different daily divides. By comparing the HBV model performance using different daily divides, Langsholt (2018) found that HBV performed slightly better when the daily divide for the meteorological and streamflow data matched. Thus, the HBV streamflow simulations used in this study are based on calibrations using daily streamflow observations matching the daily divide available for SeNorge_2018 input data, i.e. 07:00-07:00 CET. These 07:00–07:00 daily streamflow data are based on good quality streamflow observations at a sub-daily timescale. Overall, the Norwegian streamflow series are longer and more extensively quality controlled for the traditional daily data (00:00–00:00 CET). As the LSTM model is not restricted by process-based expressions, and typically perform better the more data are used, we kept the original daily meteorological and streamflow data with a mismatch in daily divides for the LSTM model, ensuring good quality data also for the 97 catchments not included in the reference catchments. The HBV and LSTM models were evaluated against observational series matching their respective daily divide, to ensure the fairest evaluation for both models with regards to flood peak timing and magnitudes. Because the two daily observational streamflow series share 17 hours, they have a high degree of correspondence (catchment-average NSE of 0.97). However, flood peaks magnitudes are typically somewhat different, and for floods occurring during the night, peaks may be allocated to two different days. Thus, we applied the same daily divide during training/calibration as during evaluation of each model.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

2.4 Long Short-Term Memory networks

Recurrent neural networks operate on time series data by processing a d-dimensional input sequence $\{x_t \in \mathbb{R}^d : t = 1, 2, \ldots, T\}$ and generating a p-dimensional output sequence $\{y_t \in \mathbb{R}^p : t = 1, 2, \ldots, T\}$ through recursive computations. At each time step t, the network updates its internal state and produces an output according to:

$$h_t = f(x_t, h_{t-1}), \quad y_t = g(h_t), \tag{1}$$

where $h_t \in \mathbb{R}^k$ represents the k-dimensional hidden state that encodes information from previous time steps, while f and g are parametrised functions that define the network's dynamics and output mapping, respectively.

LSTM networks implement a sophisticated version of the state update function f that allows the LSTM to selectively retain or forget information over long sequences, mitigating the vanishing gradient problem that affects traditional recurrent neural networks (Gers et al., 2000). In the context of streamflow prediction, this long-term memory capability enables the model to represent snowpack dynamics by retaining information about precipitation and temperature conditions during winter months, then utilizing this stored information to predict the timing and magnitude of snowmelt-driven streamflow in spring and summer (Lees et al., 2022). For a detailed decomposition of the LSTM state update function, see Kratzert et al. (2018).

To predict streamflow at day t, the model requires an input sequence $x_{(t-m):t} = \{x_\tau : \tau = t-m, t-m+1, \ldots, t\}$ of length $m+1$, which is processed through the LSTM to produce a corresponding output sequence $y_{(t-m):t}$. Our predictor, $\hat{q}_t$, the streamflow at day t, is obtained by taking the final element of this output sequence, i.e., $\hat{q}_t = y_t$. To simulate streamflow for the next timestep, the input window is shifted one time step forward at each prediction.

To train the LSTM model and estimate the parameters a, b, and all parameters within the state update function f, we employed the catchment-averaged Nash-Sutcliffe efficiency (NSE) as our loss function in line with best practices discussed in Kratzert et al. (2019b). The NSE loss function is minimised using the Adam optimiser (Kingma and Ba, 2017) and we use a linear output activation function. To mitigate overfitting during training, the final hidden state vector h_t is passed through a dropout layer, which randomly sets each element to zero with probability p_{dropout}. The dropout probability, mini-batch size (the number of training examples used to compute the loss gradient in each model update), hidden state dimension k, input sequence length m, and learning rate were treated as hyperparameters and tuned as described in Sect. 2.5.

2.5 Training the LSTM model

We implemented our LSTM model using the `cudalstm` architecture from the `neuralhydrology` package (Kratzert et al., 2022), a Python library containing a collection of LSTM-based models for hydrological modeling built on the PyTorch framework. The `cudalstm` model provides an efficient GPU-accelerated implementation of the LSTM architecture described in Sect. 2.4.

The LSTM model was trained using the 15-years period 2004–2018, to match the calibration period used for our benchmark model simulations. A selection of hyperparameters was tuned in a 5-fold cross-validation (CV) over the period 2004–2018. For each of the five CV-iterations, the period was split into a validation period of three consecutive years, and a left-out-fold of three

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

consecutive years. Each year was used as left-out-fold once and validation period once during the CV. The years adjacent to the left-out fold were removed to have a temporal 'buffer' between the left-out-fold and the 'seen' years, and the remaining years comprised the training period in each CV-iteration. The following hyperparameter values were considered in a grid search in each CV split (selected values for final model in **bold**):

- Hidden state dimension = [**64**, 128, 256]
- Dropout probability = [0.2, 0.3, **0.4**]
- Batch size = [**64**, 128, 256]
- Input sequence length = [**270**, 365]
- Learning rate = [**0: 1e-3 10: 5e-4 25: 1e-4**, 0: 5e-3 10: 1e-3 25: 5e-4]

where the learning rate schedule notation indicates the learning rate value at specified epochs (i.e. epochs 0, 10, and 25). This yielded 108 hyperparameter combinations per cross-validation split. The number of training epochs (from 1 to 50) was selected by evaluating validation NSE at each epoch for each hyperparameter combination.

Models using the faster learning rate schedule frequently exhibited poor performance (low or negative validation NSE) and training instability, leading us to exclude these configurations from further consideration. For the remaining 54 hyperparameter combinations per split, we evaluated performance on the left-out fold using the three epochs with highest validation NSE: epoch 26 (NSE=0.840), epoch 35 (NSE=0.839), and epoch 37 (NSE=0.836).

The final hyperparameter configuration (shown in bold above) was selected as the combination yielding the highest average NSE across all left-out folds, which was consistent across all three evaluated epochs. The final model was trained using 2004–2013 as the training period and 2014–2018 as the validation period. We selected epoch 37 for the final model as it achieved one of the two highest validation NSE values (0.842) and was among the three best-performing epochs identified in the cross-validation.

Several model configuration choices were fixed without hyperparameter tuning. The forget gate bias was initialised to 3, a common practice that encourages the LSTM to retain information early in training. To stabilise training, gradient norms were clipped to a maximum value of 1. The model was validated at every epoch, evaluating performance on 50 randomly selected catchments from the validation set.

The model was trained using four meteorological input time series for each catchment: catchment-averaged daily precipitation sum (mm day^{-1}), catchment-averaged mean daily temperature (°C), catchment-minimum mean daily temperature (°C), and catchment-maximum mean daily temperature (°C). In addition to these dynamic inputs, the model incorporated 21 static catchment attributes as listed in Table 1. Following best practice (Kratzert et al., 2024, 2019b), a single LSTM model was trained across all catchments, enabling the model to learn general hydrological relationships from the diverse set of catchments.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

2.6 Detection of flood events and flood generating processes

Flood events were selected from the observational time series based on a peak over threshold (POT) approach, following the methods described in Vormoor et al. (2016). Daily streamflow values exceeding the 98th percentile calculated from the 30-year period 1994–2023 were used to identify the flood peaks. To ensure independence between flood events, only the maximum streamflow value was selected within a catchment-specific time window, using the python library `pyextremes` (Bocharov, 2023). This time window was set to twice the normal flood duration (NFD), which is defined as the sum of concentration time plus recession time of the largest flood events. The full description of the estimation of NFD is provided in (Vormoor et al., 2016). In general, a concentration time of 2–3 days was found for all catchments subject to an HBV modelling experiment inputting twice the annual rainfall over a two-day period to a fully drained catchment. The median of two days was then chosen for all catchments. Recession times were estimated using the methods described in Skaugen and Onof (2014), which fit a recession curve to the largest flood events from the observational data series. Our 200 gauging stations have NDF spanning 2 to 45 days.

Flood generating process (FGP) was defined for each event based on the relative contribution of rainfall and snowmelt to the flood, based on the approach in Vormoor et al. (2015). To account for antecedent conditions, the contribution to a specific flood was computed as the sum of rainfall and snowmelt in a catchment-specific time window prior to, and including, the flood peak day. This time window was set as the catchment-specific recession times as defined above. Figure 2 shows the average fractional contribution of rainfall to flood events, reflecting the dominant flood generating process (FGP) in each catchment. Finally, all flood events were classified into snowmelt generated events, rainfall generated events and mixed events depending on whether more than two thirds of the contribution stemmed from snowmelt, rainfall or neither, respectively.

2.7 Model evaluation

Both LSTM and our benchmark HBV model were evaluated over 103 catchments, using a period unseen during model training/calibration, i.e. 15 years comprising 1994–2003 and 2019–2023. We did not extend the evaluation period further back, because early 1990s is typically considered the start of thoroughly quality-controlled rating curves for a large proportion of the Norwegian streamflow gauging stations. Additionally, older periods are more uncertain with regards to the SeNorge_2018 precipitation estimates due to lower spatial coverage of the measurement network, which is a particular problem for the HBV model due to the mass balance constraint. We did not want the HBV model to suffer from changes in the water balance from the calibration period to the evaluation period due to changes in the precipitation measuring network. To ensure this was not the case, we computed the runoff coefficient, defined as the ratio of streamflow to precipitation for the two periods, and found overall consistent results (Fig. A1). Negative discharge values in the LSTM simulations were truncated to zero prior to the evaluation, and simulated and observed series for each catchment were masked by data gaps existing in any of those series.

We evaluated model performance in terms of (i) per-catchment overall performance, (ii) descriptive statistics over all collected peaks of each flood type, and (iii) per-catchment flood peak performance of each flood type. In terms of (i), we assessed the fit between observed and simulated streamflow across the evaluation period for each catchment using Nash-Sutcliffe effi-

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

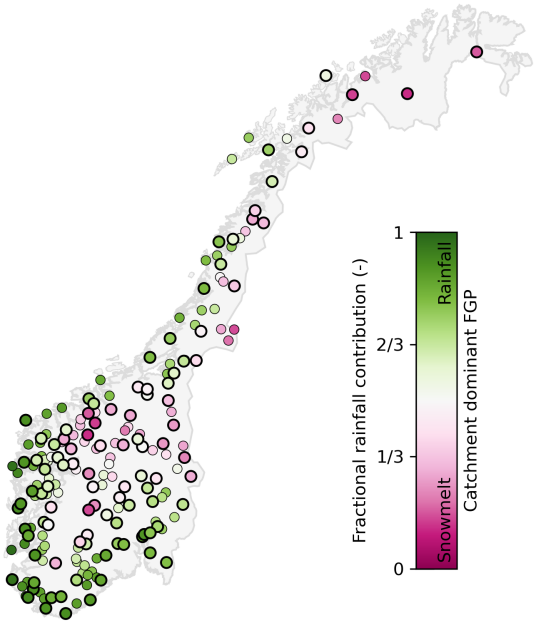

Figure 2. Average fractional rainfall (i.e. proportion of total) contribution to all observed flood events in the period 1994–2023, and its relation to the catchment dominant flood generating process (FGP). Dark pink (green) dots represent catchments with snowmelt (rainfall) as their dominant FGP, whereas near-white dots represent catchments with similar average contributions from rainfall and snowmelt. All 200 catchments used for training the LSTM are shown, with thicker circles representing the 103 catchments used for model evaluation.

ciency (NSE), the updated Kling-Gupta efficiency (KGE; Kling et al., 2012), and mean absolute error (MAE), all defined in Table 2. Additionally, we assessed the three components comprising KGE: Pearson correlation coefficient (ρ), bias ratio (β) and variability ratio (γ). The two latter are defined as:

$$\beta = \frac{\bar{\hat{q}}}{\bar{q}} \tag{2}$$

and

$$\gamma = \frac{\sigma_{\hat{q}}/\bar{\hat{q}}}{\sigma_q/\bar{q}} \tag{3}$$

where $\bar{\hat{q}}$ is mean simulated streamflow, $\bar{q}$ is mean observed streamflow, $\sigma_{\hat{q}}$ is standard deviation of simulated streamflow, and σ_q is standard deviation of observed streamflow.

In step (ii), we used descriptive statistics to summarise the flood peak performance in terms of timing and magnitudes over all collected flood peaks (i.e. independent of catchment) of each flood type: snowmelt generated floods, mixed floods and rainfall generated floods. Timing was evaluated by computing the percentage of peaks of a given flood type simulated

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

at the day of observed peak (i.e. correct timing), one day too early or too late, or more than one day too early or too late. Flood peak magnitudes were evaluated by percent errors, considering both the simulated discharge magnitude at the day of observed flood peak ($\text{PE}_{\text{peak,pinned}}$) and the simulated maximum discharge within a two-day time window of the observed flood peak ($\text{PE}_{\text{peak,floating}}$). The latter allows comparison of peak magnitudes even if the model does not match the exact day of the observed peak. The definitions of these metrics are as follows. For each catchment $s \in \mathbf{S}$, let n be the number of observed floods of a specific type at catchment s. Let t_j denote the timestep of the jth observed flood peak. Define $q_{s,j} := q_{s,t_j}$ and $\hat{q}_{s,j} := \hat{q}_{s,t_j}$ as the observed and simulated streamflow at the jth peak, respectively. Then, $\text{PE}_{\text{peak,pinned}}$ for peak j at catchment s is

$$\text{PE}^{s,j}_{\text{peak,pinned}} = \frac{\hat{q}_{s,j} - q_{s,j}}{q_{s,j}} \cdot 100, \tag{4}$$

for $j = 1, \ldots, n$ and $s \in \mathbf{S}$. Further, defining the window $\mathbf{I}_j = \{i \in \mathbb{Z} : |i - t_j| \leq 2\}$, $\text{PE}_{\text{peak,floating}}$ is

$$\text{PE}^{s,j}_{\text{peak,floating}} = \frac{\max_{i \in \mathbf{I}_j} \hat{q}_{s,i} - q_{s,j}}{q_{s,j}} \cdot 100. \tag{5}$$

In the final evaluation step (iii), we evaluated per-catchment timing error and mean absolute percent error of flood peaks separately for each flood type, as defined in Table 2. Timing error represents the percent of flood peaks that are simulated at a different day than the observed flood peaks. Corresponding to the definitions of $\text{PE}_{\text{peak,pinned}}$ and $\text{PE}_{\text{peak,floating}}$ above, we computed the mean absolute percent error of flood peaks considering both pinned ($\text{MAPE}_{\text{peak,pinned}}$) and floating ($\text{MAPE}_{\text{peak,floating}}$) simulated peaks. The per-catchment flood peak performance metrics were only computed for catchments with minimum ten flood events of the given flood type in the evaluation period. A minimum of ten flood events during the evaluation period were found for 46 catchments for snowmelt generated floods, 58 catchments for rainfall generated floods (including 13 of the same catchments that had minimum ten snowmelt generated floods), and two catchments for mixed floods. Due to the low number of catchments with the required number of mixed flood events, the results for mixed floods are excluded from the per-catchment flood peak performance results.

3 Results

3.1 Overall model performance

The LSTM model performance in terms of NSE and KGE are shown for the 103 evaluated catchments in Fig. 3. Importantly, all results were computed for the unseen evaluation period (ref. Sect. 2.7). Average scores were 0.85 for NSE and 0.87 for KGE, and all scores exceeded 0.7. The KGE components (ρ, β and γ) and MAE results are presented in Fig A2. All ρ (i.e. correlation coefficient) scores exceeded 0.84, whereof 83 % of them exceeded 0.9. A total of 91 % of the catchments had a β (i.e. bias ratio) in the range 0.9-1.1 with a similar percentage of catchments below (40 %) and above (60 %) the optimum of 1. For γ (variability ratio), 70 % of the catchments had values in the range 0.9 to 1.1, and in 90 % of the catchments the coefficient of variation was underestimated (i.e. γ<1). MAE scores ranged 0.2 to 4.3 mm day^{-1} with a average of 1.1 mm day^{-1}.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

Table 2. Definitions of per-catchment evaluation metrics used in this study: Nash-Sutcliff efficiency (NSE), Kling-Gupta efficiency (KGE), mean absolute error (MAE), mean absolute percent error ($\text{MAPE}_{\text{peak,pinned}}$ and $\text{MAPE}_{\text{peak,floating}}$) and timing error. Here q_t and $\hat{q}_t$ are observed and simulated streamflow for a specific catchment at timestep t, $\bar{q} = \frac{1}{T}\sum_{t=1}^{T} q_t$ is the mean observed streamflow, $\mathbb{I}(\cdot)$ is the indicator function and $d_j = \hat{t}_j - t_j$ is the difference in days between the simulated and observed peak for event j $(j = 1, \ldots, n)$, where n is the number of observed floods of a specific type at that catchment. We used adjusted KGE such that ρ is the Pearson correlation coefficient, β is the bias ratio (Eq. 2), and γ is the variability ratio (Eq. 3), all computed over the period $t = 1, \ldots, T$. The quantities $\text{PE}^{j}_{peak,floating}$ and $\text{PE}^{j}_{peak,pinned}$ are defined in Eq. 4 and Eq. 5, respectively.

Evaluation metric			Unit	Range	Optimum
Per-catchment overall performance (over the full time series)					
NSE	Nash-Sutcliffe efficiency	$1 - \frac{\sum_{t=1}^{T}(q_t - \hat{q}_t)^2}{\sum_{t=1}^{T}(q_t - \bar{q})^2}$	-	$(-\infty, 1]$	1
KGE	Kling-Gupta efficiency	$1 - \sqrt{(\rho-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$	-	$(-\infty, 1]$	1
MAE	Mean absolute error	$\frac{1}{T}\sum_{t=1}^{T} \lvert q_t - \hat{q}_t \rvert$	mm day^{-1}	$[0, \infty)$	0
Per-catchment flood peak performance for a specific flood type					
Timing error	Percent flood peaks simulated at a different day than the corresponding observed flood peaks	$\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(d_j \neq 0) \cdot 100$	%	$[0, 100]$	0
$\text{MAPE}_{\text{peak,pinned}}$	Mean absolute percent error of simulated flood peak magnitude at the day of observed flood peak ('pinned')	$\frac{1}{n}\sum_{j=1}^{n} \left\lvert \text{PE}^{j}_{\text{peak,pinned}} \right\rvert$	%	$[0, \infty)$	0
$\text{MAPE}_{\text{peak,floating}}$	Mean absolute percent error of simulated maximum discharge within a two-day time window of the observed flood peak ('floating')	$\frac{1}{n}\sum_{j=1}^{n} \left\lvert \text{PE}^{j}_{\text{peak,floating}} \right\rvert$	%	$[0, \infty)$	0

To benchmark our results, the LSTM NSE and KGE values were compared with those of the HBV model (Fig. 4). A total of 98 catchments (95 %) had a higher NSE score for LSTM as compared to HBV, with the score difference exceeding 0.1 for 22 of the catchments. In terms of KGE, a higher LSTM score was found for 74 of the 103 catchments (72 %), with a KGE difference exceeding 0.1 for 11 catchments. None of the HBV scores exceeded the corresponding LSTM scores by more than 0.1. The overweight of higher LSTM scores is reflected in the overweight of catchments above the diagonal line in Fig. 4c,d. The largest improvement in scores when comparing LSTM to HBV were found for catchments with HBV scores below 0.7. Most catchments with floods predominantly generated by snowmelt (i.e. pink coloured dots) had NSE and KGE scores exceeding 0.8 for both models. A mix of rainfall and snowmelt generated catchments are located both above and below the diagonal line, with no consistent pattern in the best performing model and dominant flood generating process. In terms of the KGE components, LSTM had a better score for ρ in 95 %, β in 60 %, and γ in 34 % of the catchments. In all but three catchments, mean absolute errors were lower (i.e. better) for LSTM as compared to HBV.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

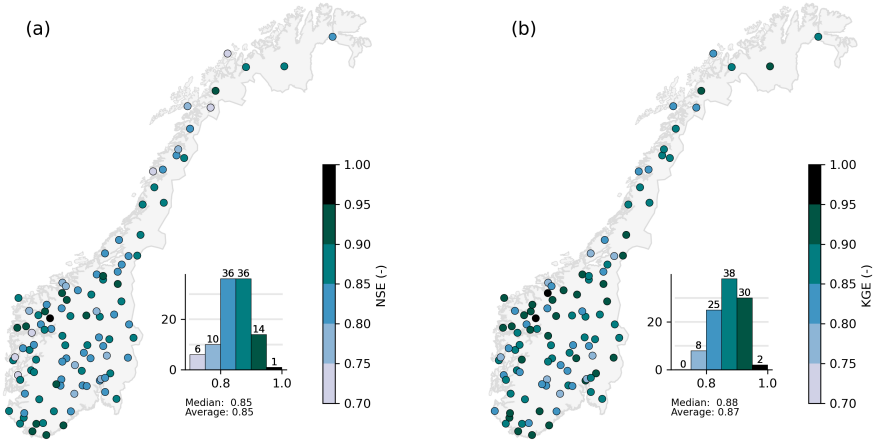

Figure 3. (a) Nash Sutcliffe efficiency (NSE) and (b) Kling-Gupta efficiency (KGE) of LSTM for the 103 evaluated catchments.

3.2 Model performance over all peaks of each flood type

Figure 5 shows the overall ability of LSTM and HBV to simulate the correct timing of the flood peaks, shown separately for snowmelt generated events, mixed events and rainfall generated events. Both models had a 24 pp higher percentage of correctly simulated peak day for rainfall generated events as compared to snowmelt generated events. LSTM simulated the correct peak day for 53 % of the snowmelt generated events and 77 % of the rainfall generated event, both numbers being 13 pp higher than the corresponding numbers for HBV. For mixed events, the model difference was smaller, with 65 % (LSTM) and 62 % (HBV) of the events simulated at the correct peak day. For all flood types, a higher percentage of peak events were simulated too early by HBV than by LSTM. For example, 36 % of the snowmelt generated events are simulated too early by HBV, as compared to 21 % for LSTM. For rainfall generated events, the corresponding numbers are 17 % for HBV and 9 % for LSTM.

The distributions of percent error of simulated flood peak magnitudes are relatively similar for snowmelt, mixed and rainfall generated events (Fig. 6). Slightly better results were found for $PE_{peak,floating}$ as compared to $PE_{peak,pinned}$, as expected since the former allows for extracting the simulated flood peak despite the model missing the peak day by one or two days (ref. Fig. 5). For LSTM, close to half of the snowmelt (49 %) and rainfall generated events (48 %) are within ±20 % of $PE_{peak,floating}$ values, compared to 44 %, respectively 43 %, of the events for HBV. Regardless of model and flood type, most (76 to 87 %) of the observed peak magnitudes were underestimated. More events had relatively large underestimations (-40 to -80 %) for HBV as compared to LSTM.

3.3 Per-catchment flood peak performance

The resulting per-catchment flood peak performance scores for LSTM and HBV are shown in Fig. 7, and maps showing the differences in scores between the two models are presented in Fig. 8. Across all peak metrics and flood types, LSTM had

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

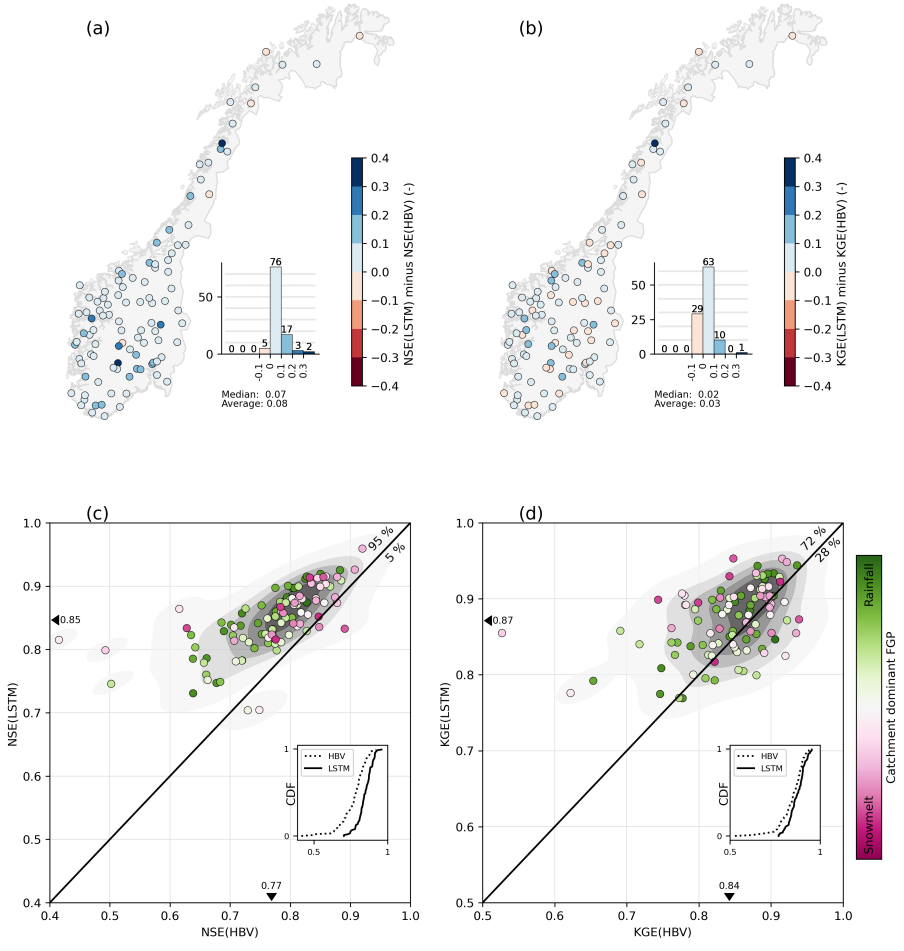

Figure 4. Difference between LSTM and HBV in (a) NSE and (b) KGE. Blue color implies a higher score for LSTM for that catchment, whereas red color implies a higher score for HBV. Scatterplots show (c) NSE and (d) KGE for HBV (x-axis) versus LSTM (y-axis). Background is shaded by the corresponding 2D empirical density function. Each dot represents a catchment, and dots above the diagonal line represent higher scores for LSTM as compared to HBV. Triangles with adjacent numbers represent catchment average score for each model. Each catchment is coloured by the corresponding catchment's dominant flood generating process (FGP, ref. Fig. 2). Inserted in the lower left corner of (c) and (d) are the cumulative density functions (CDFs) of the two models' NSEs and KGEs, respectively.

310 better scores than HBV for the majority of the catchments. LSTM had the lowest percentage of timing error in 64 % of the catchments for snowmelt generated flood events, and 83 % of the catchments in terms of rainfall generated flood events. Median improvement by LSTM across catchments was 14 pp for snowmelt and 15 pp for rainfall generated floods, but improvements exceeding 24 pp were found for several catchments. Both models had a markedly lower timing error when considering rainfall generated events (catchment averages of 24 % for LSTM and 37 % for HBV) as compared to snowmelt generated events
315 (catchment averages of 47 % for LSTM and 61 % for HBV).

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

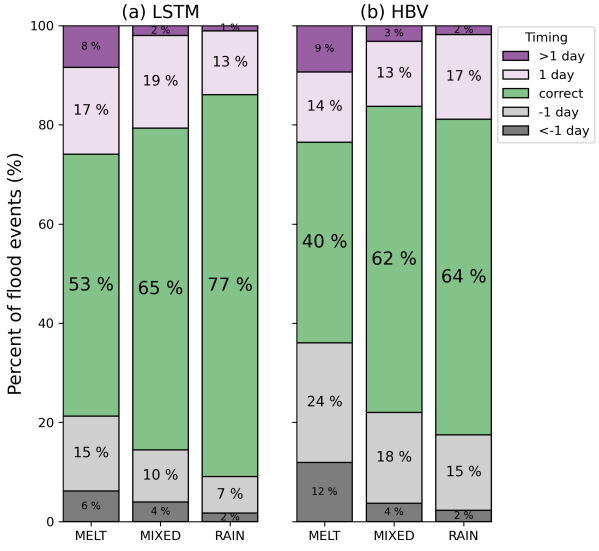

Figure 5. Flood peak timing results for (a) LSTM and (b) HBV of all collected snowmelt generated floods ('MELT'), mixed floods ('MIXED') and rainfall generated floods ('RAIN'). Shown are the percentages of flood events where the peak day was simulated at the correct day (green), one day too early (light grey), even earlier (dark grey), one day too late (light purple) or even later (dark purple) as compared to observed flood peak day.

Most $MAPE_{peak,pinned}$ scores were in the range 10 to 40 %, with catchment average of 24 % to 25 % for LSTM and 29 % for HBV. $MAPE_{peak,pinned}$ for snowmelt generated floods had the largest median model difference of 6 pp, in the favour of LSTM. For both error metrics considering peak magnitude ($MAPE_{peak,pinned}$ and $MAPE_{peak,floating}$), LSTM had smaller errors than HBV for a larger proportion of the catchments when considering snowmelt generated floods as compared to rainfall generated floods. On the other hand, more catchments had a relatively larger difference between the models when considering rainfall generated floods, in particular for catchments where HBV have a relatively high (>40 %) error score.

Figure 9 combines the model comparisons of overall performance (KGE) and flood peak performance ($MAPE_{peak,pinned}$). Catchments in quadrant IV (lower right square in each plot) have a better metric score for LSTM as compared to HBV in terms of both KGE and $MAPE_{peak,pinned}$. The majority of catchments (57 and 51 %) are found in quadrant IV, including the majority of catchments with snowmelt as dominant FGP. A few catchments show notably better results for LSTM, with a >0.08 larger KGE and a smaller $MAPE_{peak,pinned}$ of more than 10 pp. The minority (7 and 10 %) of the catchments are found in quadrant II that represents better HBV scores for both metrics. Figures A3 and A4 show model difference in NSE, respectively KGE, versus model difference in all three flood peak metrics. In the corresponding plots using NSE instead of KGE, nearly all catchments are located in quadrants I and IV, as LSTM had the highest NSE scores for 95 % of the catchments.

A per-catchment summary of the best-performing model in terms of all evaluated metrics is presented in Fig. 10. LSTM had the best scores for the majority of the catchments for all evaluated metrics except (the KGE component) γ. For MAE, NSE and ρ, 95 to 97 % of the catchments had a better score for LSTM. To indicate the catchment specific best model across metrics,

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

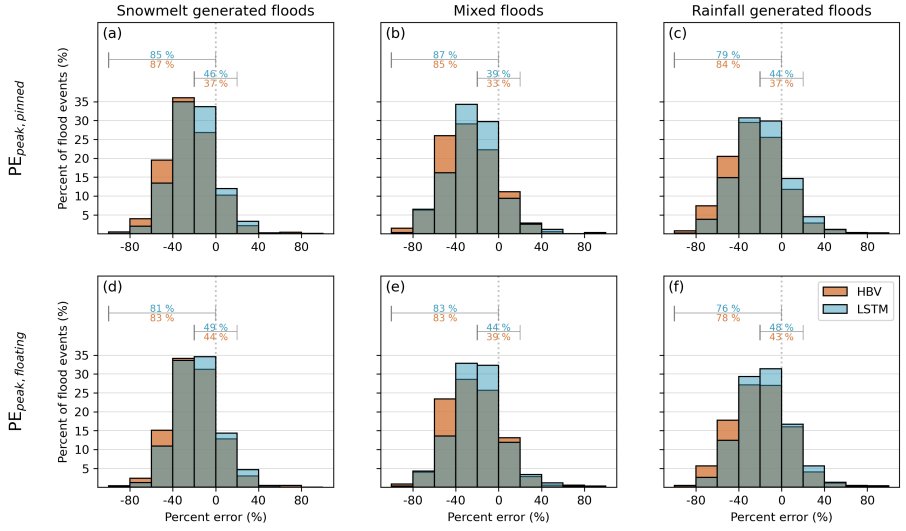

Figure 6. Percent error (PE) of flood peak magnitudes simulated by HBV (orange) and LSTM (blue) for the three different types of flood events: snowmelt generated floods (a,d), mixed floods (b,e) and rainfall generated floods (c,f). Upper panel (a–c) shows the pinned simulations ($PE_{peak,pinned}$ comparing discharge magnitudes at the day of observed peak), whereas lower panel (d–f) shows the floating simulations ($PE_{peak,floating}$ comparing observed peak discharge with maximum simulated discharge within two days of observed peak). In each subplot, the percentage of events that were underestimated (numbers at horizontal line extending -100 to 0 %), and the percentage of events with a relative error within ±20 % (numbers at horizontal line extending -20 to 20 %) are shown for LSTM (blue) and HBV (orange).

the percentage of metrics where LSTM had a better score is given at the top of each catchment column. For five of the 103 catchments, HBV had better scores for minimum two thirds of the metrics. LSTM on the other hand, had better scores for minimum two thirds of the metrics for 80 of the catchments.

4 Discussion

The LSTM model demonstrated good performance across the metrics assessed in the study. As compared to the benchmark model (HBV), performances were improved for a majority of the catchments in terms of all but one evaluated metrics. The largest improvements by LSTM were often found for metrics and catchments where HBV scores were relatively poor. Thus, the results imply that LSTM can improve hydrological services and flood specific assessments in regions affected by both snowmelt and rainfall generated floods.

4.1 LSTM's ability to simulate different aspects of the streamflow series

LSTM NSE scores were generally high, exceeding 0.7 for all 103 catchments, and with an average NSE of 0.85 as compared to 0.77 for HBV. Many of the snowmelt dominated catchments were among the catchments with the highest NSE scores for both

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

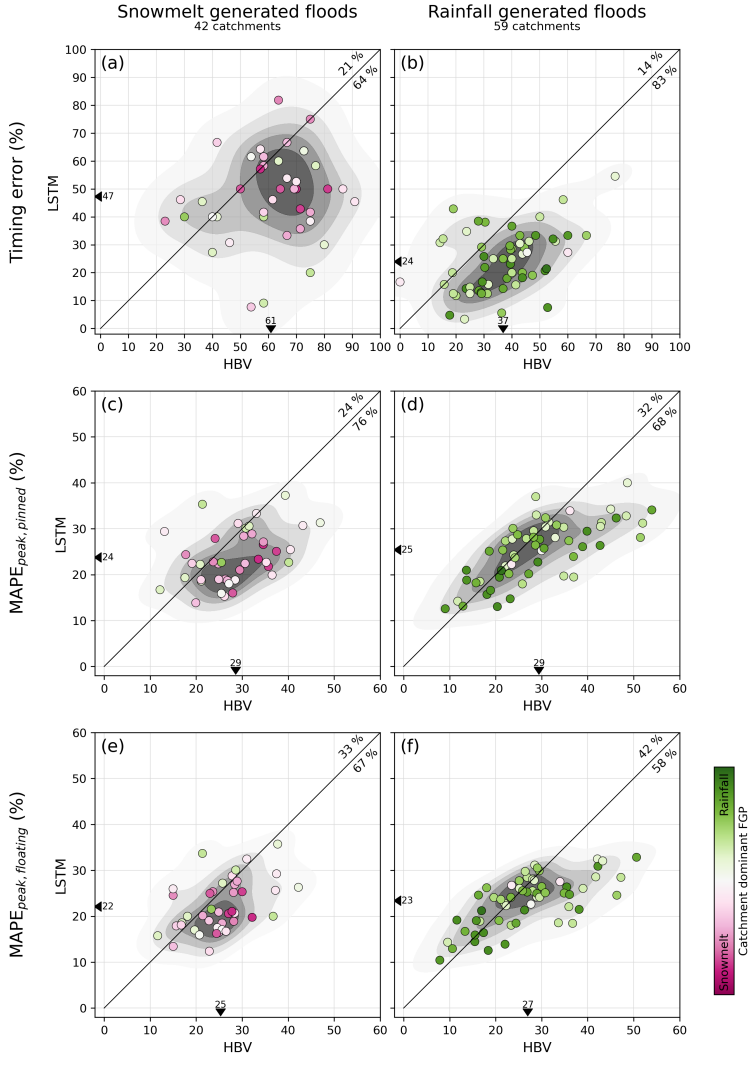

Figure 7. Flood peak performance scores per catchment for HBV (x-axis) versus LSTM (y-axis), for snowmelt generated floods (left) and rainfall generated floods (right). Upper panel (a–b) shows the peak timing error (percentages of incorrectly simulated days of peak discharges). Middle panel (c–d) and lower panel (e–f) show the mean absolute percent error for pinned simulations ($MAPE_{peak,pinned}$ comparing discharge magnitudes at the day of observed peak), and for floating simulations ($MAPE_{peak,floating}$ comparing observed peak discharge with maximum simulated discharge within two days of observed peak), respectively. Background is shaded by the corresponding 2D empirical density function. Triangles with adjacent numbers represent catchment average score for each model. Dots below the diagonal line represent catchments with better (i.e. smaller error) metric values for LSTM as compared to HBV. The percentages of catchments above and below the diagonal lines are provided in the upper right corner of each scatterplot. Each catchment is coloured by the corresponding catchment's dominant flood generating process (FGP, ref. Fig. 2).

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

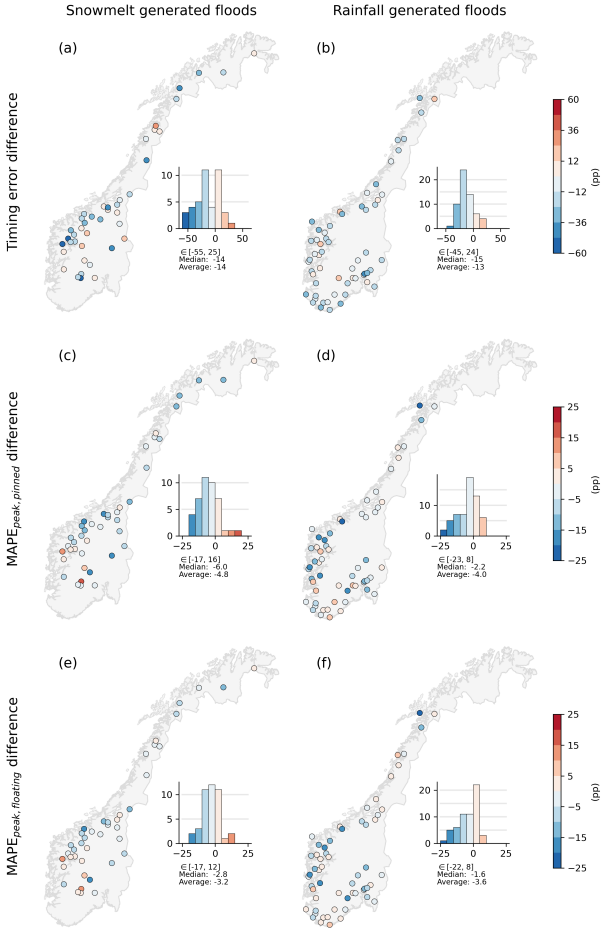

Figure 8. Model difference (LSTM minus HBV results) in per-catchment flood peak performance scores for snowmelt generated floods (left) and rainfall generated floods (right). Upper panel (a–b) shows the model difference in timing error. Middle panel (c–d) and lower panel (e–f) show the model difference in mean absolute percent error for pinned simulations ($MAPE_{peak,pinned}$ comparing discharge magnitudes at the day of observed peak), and for floating simulations ($MAPE_{peak,floating}$ comparing observed peak discharge with maximum simulated discharge within two days of observed peak), respectively. Blue colour represents a smaller error for LSTM as compared to HBV. Alongside each map, the corresponding histogram, range, median and average difference are provided.

models, in line with previous studies (Anderson and Radić, 2022; Ruzzante et al., 2025). The higher percentage of catchments improved by LSTM in terms of NSE (95 %) as compared to KGE (72 %), relates to the use of NSE as loss function in the training of LSTM.

Of the three components constituting KGE, the one reflecting the variability (γ) of the model was the only evaluated metric with a better HBV score for the majority of the catchments (Fig. 10). LSTM generally underestimated γ for more catchments and to a larger degree than HBV (Fig. A2). The underestimated variability may partly relate to the applied loss function (NSE)

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

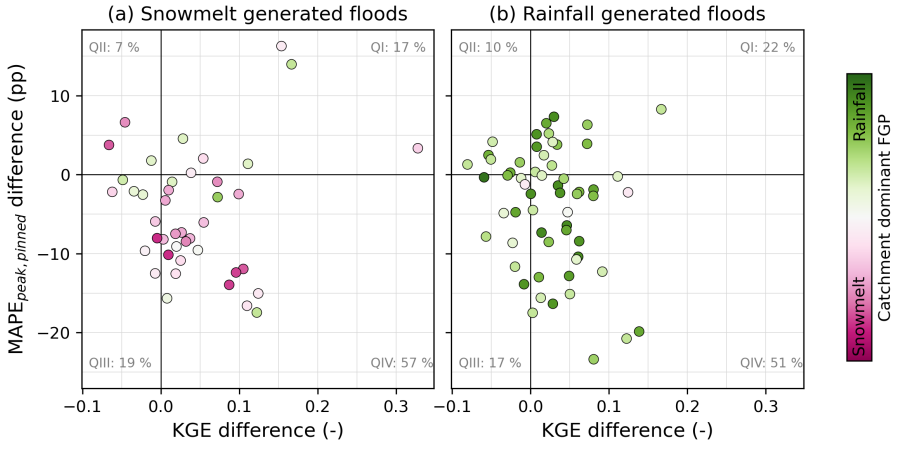

Figure 9. Model difference (i.e. LSTM minus HBV) in the overall performance metric KGE (x-axis) versus the peak metric $MAPE_{peak,pinned}$ for (a) snowmelt generated floods and (b) rainfall generated floods. Each dot represents a catchment and is coloured by the catchment dominant flood generating process (FGP; ref. Fig. 2). Catchments within quadrant IV (QIV; lower right square) have a better score for LSTM as compared to HBV both in terms of overall score and peak metric score. Percentages of catchments within each quadrant are given in the corners of the plots.

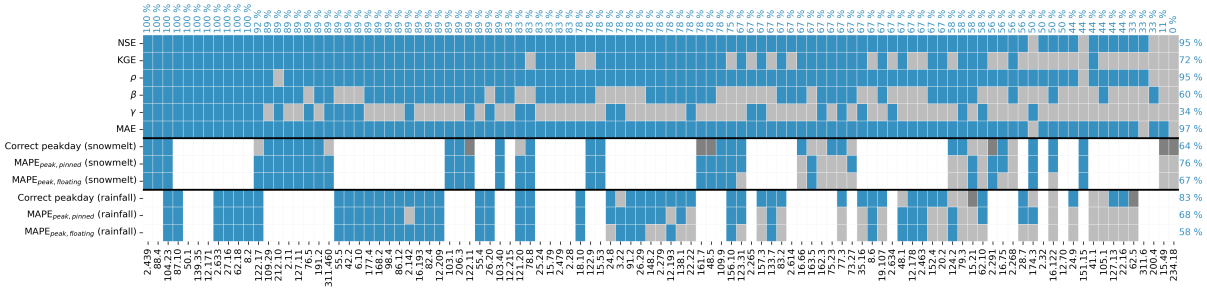

Figure 10. Summary of best performing model in terms of all evaluated metrics and catchments. LSTM has the best score for blue cells, HBV has the best score for light grey cells, and both models have equal scores for dark grey cells. Percentages of blue cells for each catchment (i.e. column) and metric (i.e. row) are provided at the top and right, respectively.

which normalises the prediction errors by the variability of the time series. Gupta et al. (2009) has previously showed how the variability has to be underestimated to maximise NSE. The two models' results for the KGE component reflecting water balance (β) were similar (Fig. A2). A slight majority of catchments exhibited a better β for LSTM than HBV despite that no water balance constraints were implemented in the LSTM model as opposed to HBV. Further, LSTM outperformed HBV in all but two catchments in terms of mean absolute error, implying a systematically higher accuracy in the LSTM predictions.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

4.2 Within-model differences in flood peak performance for snowmelt versus rainfall generated floods

The results revealed notable differences in the ability of LSTM to simulate the correct peak timing and magnitude of snowmelt versus rainfall generated floods. Whereas 77 % of rainfall generated flood peaks were simulated at the correct day, the corresponding percentage for snowmelt generated flood peaks was only 53 % (Fig. 5). As snowmelt generated floods react to snowmelt typically spanning days to weeks, the floods often last multiple days and may not have a very distinct peak as compared to rainfall generated floods. Relatively small differences between peak discharge and discharge magnitudes in adjacent days may explain the lower percentage of correctly simulated peak days for snowmelt generated floods. This reasoning is supported by the similar mean absolute percent errors at the day of observed peak ($MAPE_{peak,pinned}$) for snowmelt and rainfall generated floods (Fig. 7).

The identified differences in model performance for floods of different flood generating processes were not unique to LSTM. Similar differences were found for HBV, which also had a 24 pp difference in correctly simulated peak day, but with 13 pp lower percentages as compared to LSTM. In terms of $MAPE_{peak,pinned}$, HBV exhibited the highest errors (exceeding 40 %) for rainfall generated floods in several catchments. Similarly high errors were not found in any catchments for LSTM. Flood-type dependent performance has also previously been demonstrated by Brunner et al. (2021), who identified differences in a model's flood peak performances in catchments of different regimes across four different process-based hydrological models.

4.3 Model evaluations reflecting the characteristics of interest

The value of a hydrological model for specific applications depends on the model performance with regards to relevant characteristics. Thus, model evaluations beyond the metrics considering the entire streamflow time series (e.g. NSE and KGE) are often necessary. Figure 9 shows that the model preferred for simulating the full streamflow time series (KGE) does not always match the one preferred for reproducing peak magnitudes ($MAPE_{peak,pinned}$). Whereas the majority of the catchments had better LSTM scores for both overall and flood peak metrics (quadrant IV), the preferred model depended on the metric for more than one third of the catchments (quadrants I and III). Notably, several catchments had a slightly (<0.025) lower KGE for LSTM than HBV, whereas the peak magnitude errors were reduced by 5 to 15 pp. In such cases, LSTM may be the preferred model for flood-specific applications despite the somewhat lower overall score.

We included three different flood peak specific evaluation metrics in this study to account for different aspects relevant for flood applications. Specifically, $MAPE_{peak}$ was evaluated both for simulated discharge at the day of observed peak ('pinned'), and for simulated maximum discharge within two days of the observed peak ('floating'). For HBV, which generally had higher timing errors, the differences between the two $MAPE_{peak}$ metrics were larger than for LSTM. However, the overweight of catchments with a better LSTM $MAPE_{peak,floating}$ score make evident that peak magnitudes were better represented by LSTM regardless of the timing errors by HBV. Generally, our flood peak evaluation results demonstrated that LSTM does a better job than HBV for most catchments. At the local (i.e. catchment) scale, however, the conclusion on the preferred model depends in some cases on the desired characteristic, and the desired characteristics depend on the application. For issuing flood warning, for example, the timing may be crucial, whereas inundation mapping depend more on magnitudes being correctly simulated.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

What are important aspects may also differ for snowmelt generated events and rainfall events, such as peak timing for flood warning of long-lasting snowmelt generated floods may be less crucial than for those of rainfall generated events.

4.4 Training LSTM on the premises of deep learning models

In this study, we constrained the forcing data and training period for LSTM to match that of our benchmark model in order to have a reasonably fair comparison. By doing so, we ensured that differences in model performances could not be attributed to differences in how informed the models were about local hydrometeorological conditions in each catchment. However, given the fundamentally different nature of the two models, the premises for training the best model are different. Accordingly, premade choices suitable of the benchmark model is not the best suitable choices for deep learning models, and there is a further potential to improve the LSTM simulations. Low-hanging fruit to achieve an LSTM model with even higher performance include expanding the training period considerably where possible and forcing the model with a wider range of atmospheric variables and datasets (Martel et al., 2025; Kratzert et al., 2021). Another possibility is finetuning of the LSTM model for individual catchments to explore the potential to improve the model further for local conditions (Kratzert et al., 2018). Such potential avenues should be explored in case an LSTM model is considered for operational use to unleash the best potential based on the premises of deep learning models.

5 Conclusions

This study evaluated the ability of LSTM to simulate floods of different flood generating processes by assessing peak timing and peak magnitude of snowmelt generated floods and rainfall generated floods separately. To evaluate LSTM's potential for operational use in snow-influenced regions, the results were compared with the operational model in the study region, HBV. Our findings can be summarised by the following answers to our research questions:

1. LSTM simulated streamflow series with higher performance than HBV for 95 % of the catchments in terms of NSE and 72 % of the catchments in terms of KGE. The largest improvements by LSTM were found for catchments with the lowest HBV performance scores.

2. LSTM simulated the correct timing of the flood peaks more often for rainfall generated events (77 %) than mixed events (65 %) and snowmelt generated events (53 %). The percentages of correctly simulated peak timing were 13 pp higher than those of HBV for both snowmelt and rainfall generated events, mainly because HBV more frequently simulated peak discharge too early.

   LSTM simulated flood peak magnitudes with a similar performance for snowmelt generated floods, mixed floods and rainfall generated floods. Percent errors were within ± 20 % for close to half of the events, a slightly better result than that of HBV. Both LSTM and HBV generally underestimated flood peak magnitudes, whereas HBV had a higher proportion of the largest underestimations (more than 40 % underestimation) as compared to LSTM for the three types of floods.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

3. A large spread among catchments was found in timing errors by LSTM for snowmelt generated floods (approx. 10 to 80 %), with an average of 47 %. Timing errors for rainfall generated floods were notably better, with an average error of 24 % and only one catchment exceeding 50 %. LSTM exhibited smaller timing errors than HBV in the majority of the catchments for both snowmelt and rainfall generated floods.

   Per-catchment LSTM results of mean absolute percent error of flood peaks were in the range 10 to 40 % for both snowmelt and rainfall generated flood events. For both types of floods, the errors were smaller for LSTM as compared to HBV for the majority of the catchments. Most model differences were within 10 pp, but LSTM improved the rainfall generated flood peak magnitude simulations with up to 23 pp for catchments with relatively poor HBV scores.

Overall, LSTM provided reliable simulations of streamflow time series and floods of different flood generating processes in catchments influenced by seasonal snow. Notable improvements were found in the timing of simulated flood peaks, and in overall and peak magnitude metrics for catchments where the benchmark model had relatively poor results. Our findings highlight LSTM's potential to improve hydrological services and flood assessments in regions prone to both snowmelt and rainfall generated floods.

Data availability. Discharge data are openly available at https://hydapi.nve.no/ (NVE, 2026a). SeNorge_2018 data are openly available at https://thredds.met.no/thredds/catalog/senorge/seNorge_2018/catalog.html (MET Norway, 2026). The snowmelt dataset is openly available at https://thredds.met.no/thredds/catalog/senorge/seNorge_snow/qsw/catalog.html (NVE and MET Norway, 2026). Prepared catchment-level data used in this study will be made available at Zenodo.

Appendix A: Additional figures

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

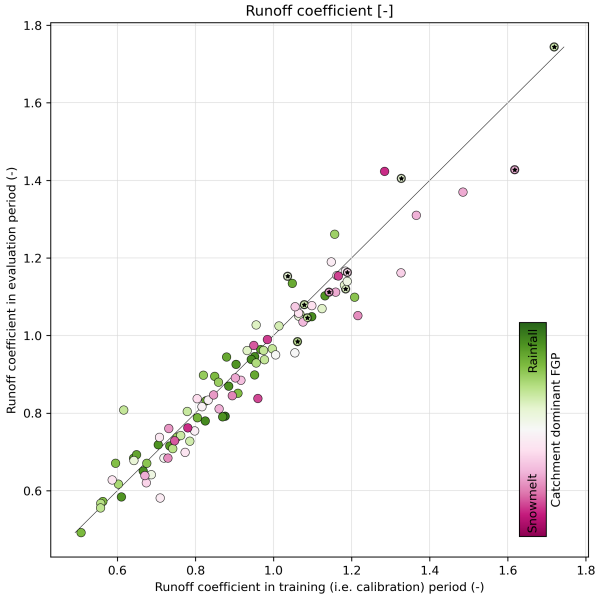

Figure A1. Runoff coefficient (i.e. ratio of streamflow (mm day^{-1}) to precipitation (mm day^{-1})) of each catchment in the training period (i.e. calibration period; x-axis) versus the evaluation period (y-axis). A dot close to the diagonal line implies consistent runoff coefficients for the two periods. Catchments with glaciers covering more than 3 % of their area are marked with stars. We note that values exceeding 1 (i.e. streamflow larger than precipitation) are mainly due to underestimation of precipitation, although glacier melt can explain part of the difference in catchments with glaciers. Each catchment is coloured by the corresponding catchment's dominant flood generating process (FGP, ref. Fig. 2).

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

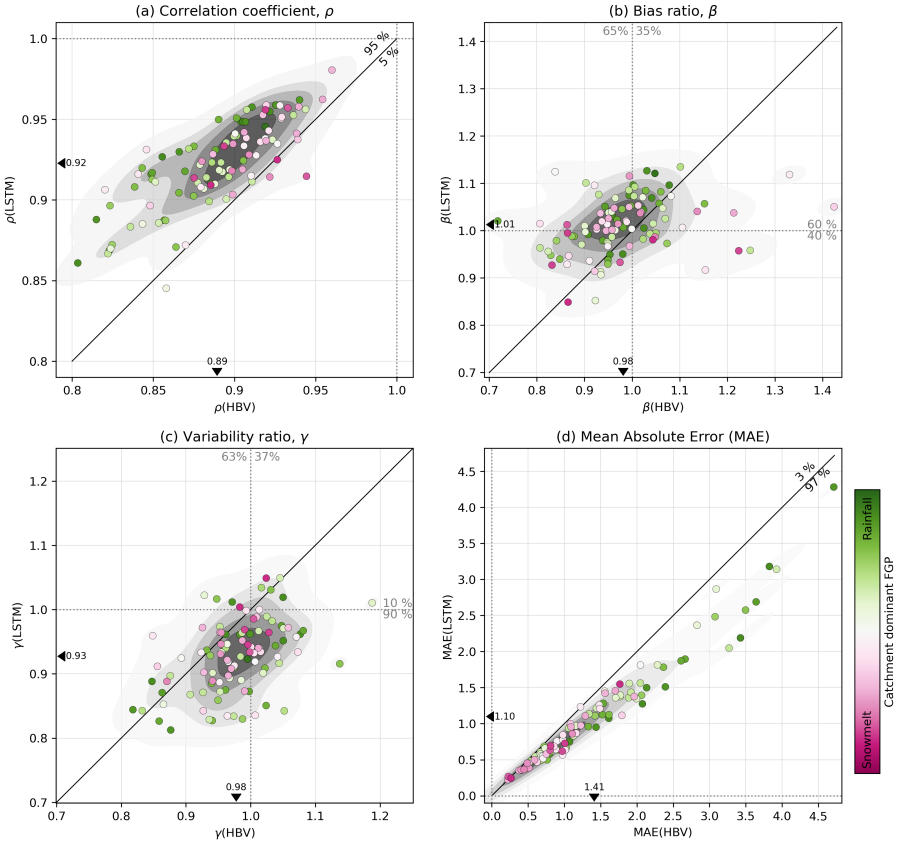

Figure A2. Overall performance per catchment for HBV (x-axis) versus LSTM (y-axis) in terms of the three KGE components (a) correlation coefficient, ρ, (b) bias ratio, β, and (c) variability ratio, γ, as well as (d) mean absolute error, MAE. Optimum values are marked with stippled lines. Backgrounds are shaded by the corresponding 2D empirical density functions. Triangles with adjacent numbers represent catchment average score for each model. The percentages of catchments above and below the diagonal lines are provided in the upper right corner of the scatterplots of ρ and MAE. For β and γ, percentages of scores below and above optimum value of 1 are provided for both models. Each catchment is coloured by the corresponding catchment's dominant flood generating process (FGP, ref. Fig. 2).

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

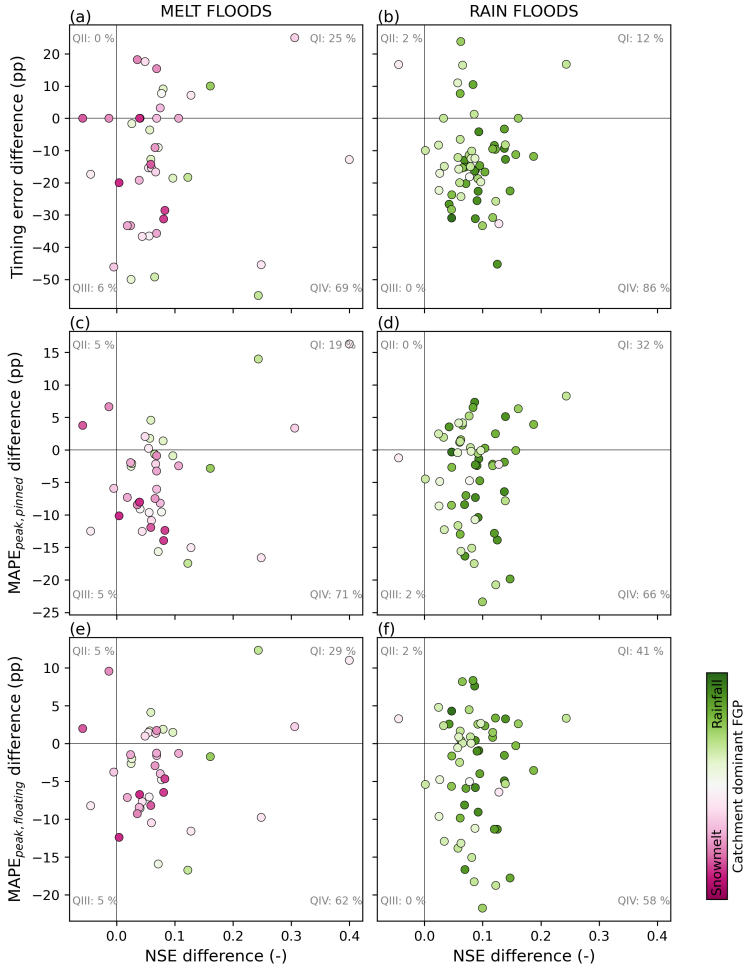

Figure A3. Model difference (i.e. LSTM minus HBV) in the overall performance metric NSE (x-axis) versus the flood peak metrics: (a–b) timing error, (c–d) $MAPE_{peak,pinned}$ and (e–f) $MAPE_{peak,floating}$, separately for snowmelt generated floods (left) and rainfall generated floods (right). Each dot represents a catchment and is coloured by the catchment's dominant flood generating process (FGP; ref. Fig. 2). Catchments within quadrant IV (QIV; lower right square) have a better score for LSTM as compared to HBV both in terms of overall score (NSE) and flood peak score. Percentages of catchments within each quadrant are given in the corners of the plots.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

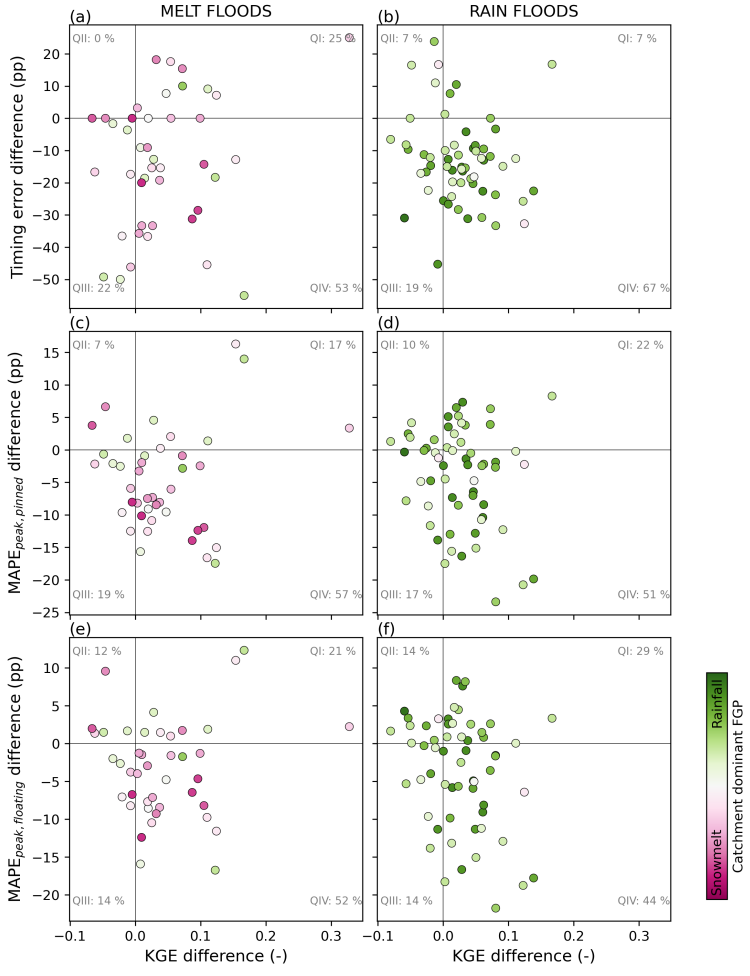

Figure A4. Model difference (i.e. LSTM minus HBV) in the overall performance metric KGE (x-axis) versus the flood peak metrics: (a–b) timing error, (c–d) $MAPE_{peak,pinned}$ and (e–f) $MAPE_{peak,floating}$, separately for snowmelt generated floods (left) and rainfall generated floods (right). Each dot represents a catchment and is coloured by the catchment's dominant flood generating process (FGP; ref. Fig. 2). Catchments within quadrant IV (QIV; lower right square) have a better score for LSTM as compared to HBV both in terms of overall score (KGE) and flood peak score. Percentages of catchments within each quadrant are given in the corners of the plots.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

Author contributions. SJB, DMB and SN designed the study with contributions from SAK and KE. All authors collected the data, and SJB carried out the data preprocessing, modelling, analyses and visualisations with contributions from DMB. SJB, with input from DMB, wrote the original draft, and all authors contributed to revision and editing of the manuscript.

Competing interests. The authors have no competing interests to declare.

Acknowledgements. Data and code providers are greatly acknowledged. We thank the Norwegian Water Resources and Energy Directorate (NVE) for providing gridded snowmelt data, catchment attributes, observed streamflow data, and simulated streamflow data from the operational HBV model. We also thank the Norwegian Meteorological institute for providing the seNorge_2018 data. We further thank the team behind neuralhydrology, who make LSTM modelling accessible for the broader hydrological community.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

References

Addor, N. and Melsen, L. A.: Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models, Water Resources Research, 55, 378–390, https://doi.org/10.1029/2018WR022958, 2019.

Anderson, S. and Radić, V.: Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling, Hydrology and Earth System Sciences, 26, 795–825, https://doi.org/10.5194/hess-26-795-2022, 2022.

Andreassen, L. M., Nagy, T., Kjøllmoen, B., and Leigh, J. R.: An inventory of Norway's glaciers and ice-marginal lakes from 2018–19 Sentinel-2 data, Journal of Glaciology, 68, 1085–1106, https://doi.org/10.1017/jog.2022.20, 2022.

Barna, D. M., Engeland, K., Kneib, T., Thorarinsdottir, T. L., and Xu, C.-Y.: Regional index flood estimation at multiple durations with generalized additive models, EGUsphere, 2023, 1–43, https://doi.org/10.5194/egusphere-2023-2335, 2023.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, Tech. rep., SMHI Report Nr RHO 7/1976. The Swedish Meteorological and Hydrological Institute (SMHI), ISSN 0347-7827, 1976.

Bocharov, G.: pyextremes, https://github.com/georgebv/pyextremes, 2023.

Brunner, M. I., Melsen, L. A., Wood, A. W., Rakovec, O., Mizukami, N., Knoben, W. J. M., and Clark, M. P.: Flood spatial coherence, triggers, and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models, Hydrology and Earth System Sciences, 25, 105–119, https://doi.org/10.5194/hess-25-105-2021, 2021.

Doherty, J.: PEST: Model Independent Parameter Estimation. Fifth edition of user manual, https://www.nrc.gov/docs/ML0923/ML092360221.pdf, 2004.

Engeland, K., Schlichting, L., Randen, F., Nordtun, K., Reitan, T., Wang, T., Holmqvist, E., Voksø, A., and Eide, V.: Utvalg og kvalitetssikring av flomdata forflomfrekvensanalyser, Tech. rep., Technical Report 85/2016. The Norwegian Water Resources and Energy Directorate (NVE), ISBN 978-82-410-1538-0, https://publikasjoner.nve.no/rapport/2016/rapport2016_85.pdf, 2016.

Engeland, K., Glad, P., Hamududu, B. H., Li, H., Reitan, T., and Stenius, S. M.: Lokal og regional flomfrekvensanalyse, Tech. rep., Technical Report 10/2020. The Norwegian Water Resources and Energy Directorate (NVE), ISBN 978-82-410-2014-8, https://publikasjoner.nve.no/rapport/2020/rapport2020_10.pdf, 2020.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrology and Earth System Sciences, 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.

GeoNorge: Totalnedbørfelt til målestasjon, https://kartkatalog.geonorge.no/metadata/totalnedboerfelt-til-maalestasjon/ac1c71db-9850-4e89-8162-2baba8b980e7, last accessed: 10.02.2026. Hosted by the Norwegian Mapping Authority (Kartverket), 2026a.

GeoNorge: ELVIS elvenett, https://kartkatalog.geonorge.no/metadata/elvis-elvenett/3f95a194-0968-4457-a500-912958de3d39, last accessed: 10.02.2026. Hosted by the Norwegian Mapping Authority (Kartverket), 2026b.

GeoNorge: Løsmasser, https://kartkatalog.geonorge.no/metadata/loesmasser/3de4ddf6-d6b8-4398-8222-f5c47791a757, last accessed: 10.02.2026. Property: infiltrasjonEvne. Hosted by the Norwegian Mapping Authority (Kartverket), 2026c.

Gers, F. A., Schmidhuber, J., and Cummins, F.: Learning to Forget: Continual Prediction with LSTM, Neural Computation, 12, 2451–2471, https://doi.org/10.1162/089976600300015015, 2000.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hagen, J. S., Hasibi, R., Leblois, E., Lawrence, D., and Sorteberg, A.: Reconstructing daily streamflow and floods from large-scale atmospheric variables with feed-forward and recurrent neural networks in high latitude climates, Hydrological Sciences Journal, 68, 412–431, https://doi.org/10.1080/02626667.2023.2165927, 2023.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Jiang, S., Bevacqua, E., and Zscheischler, J.: River flooding mechanisms and their changes in Europe revealed by explainable machine learning, Hydrology and Earth System Sciences, 26, 6339–6359, https://doi.org/10.5194/hess-26-6339-2022, 2022.

Kartverket: Høydedata, https://hoydedata.no/LaserInnsyn2/, last accessed: 10.02.2026. The Norwegian Mapping Authority (Kartverket), 2026.

Killingtveit, Å. and Saelthun, N. R.: Hydrological models, vol. 7 of *Hydropower development. Vol. 7*, pp. 99–128, Norwegian Inst. of Technology. Dept. of Hydraulic Engineering Norwegian Inst. of Technology, ISBN 978-82-7598-026-5, 1995.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://doi.org/10.48550/arXiv.1412.6980, 2017.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., and Klambauer, G.: NeuralHydrology – Interpreting LSTMs in Hydrology, pp. 347–362, Springer International Publishing, Cham, ISBN 978-3-030-28954-6, https://doi.org/10.1007/978-3-030-28954-6_19, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019b.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology — A Python library for Deep Learning research in hydrology, Journal of Open Source Software, 7, 4050, https://doi.org/10.21105/joss.04050, 2022.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Langsholt: Samkjøring av vær- og vannføringsdøgnet i hydrologiske modeller, Tech. rep., Technical Report 71/2018. The Norwegian Water Resources and Energy Directorate (NVE), ISBN 978-82-410-1728-5, https://publikasjoner.nve.no/rapport/2018/rapport2018_71.pdf, 2018.

Lawrence, D., Haddeland, I., and Langsholt, E.: Calibration of HBV hydrological models using PEST parameter estimation, Tech. rep., Technical Report 1/2009. The Norwegian Water Resources and Energy Directorate (NVE), ISBN 78-82-410-0680-7, 2009.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

Lussana, C., Tveito, O. E., Dobler, A., and Tunheim, K.: seNorge_2018, daily precipitation, and temperature datasets over Norway, Earth System Science Data, 11, 1531–1551, https://doi.org/10.5194/essd-11-1531-2019, 2019.

Martel, J.-L., Arsenault, R., Turcotte, R., Castañeda Gonzalez, M., Brissette, F., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Lachance-Cloutier, S., Rondeau-Genesse, G., and Caron, L.-P.: Exploring the ability of LSTM-based hydrological models to simulate streamflow time series for flood frequency analysis, Hydrology and Earth System Sciences, 29, 4951–4968, https://doi.org/10.5194/hess-29-4951-2025, 2025.

MET Norway: SeNorge_2018, https://thredds.met.no/thredds/catalog/senorge/seNorge_2018/catalog.html, last accessed: 11.02.2026. The Norwegian Meteorological institute (MET Norway), 2026.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., et al.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

NVE: NVE Hydrological API (HydAPI), https://hydapi.nve.no/, last accessed: 10.02.2026. The Norwegian Water Resources and Energy Directorate (NVE), 2026a.

NVE: Sildre, https://sildre.nve.no/, last accessed: 10.02.2026. The Norwegian Water Resources and Energy Directorate (NVE), 2026b.

NVE and MET Norway: Snowmelt, https://thredds.met.no/thredds/catalog/senorge/seNorge_2018/catalog.html, last accessed: 11.02.2026. The Norwegian Water Resources and Energy Directorate (NVE) and the Norwegian Meteorological institute (MET Norway), 2026.

Roksvåg, T., Vandeskog, S. M., Wulff, C., and Wergeland, K.: An LSTM network for joint modeling of streamflow and hydropower generation for run-of-river plants, Journal of Hydrology, 667, 134 890, https://doi.org/10.1016/j.jhydrol.2025.134890, 2026.

Ruan, G. and Langsholt, E.: Rekalibrering av flomvarslingas HBV-modeller med inndata fra seNorge, versjon 2.0, Tech. rep., Technical Report 71/2017. The Norwegian Water Resources and Energy Directorate (NVE), ISBN 978-82-410-1624-0, 2017.

Ruzzante, S. W., Knoben, W. J. M., Wagener, T., Gleeson, T., and Schnorbus, M.: Technical Note: High Nash Sutcliffe Efficiencies conceal poor simulations of interannual variance in tropical, alpine, and polar catchments, EGUsphere, 2025, 1–27, https://doi.org/10.5194/egusphere-2025-3851, 2025.

Sælthun, N. R.: The Nordic HBV Model, Tech. rep., Technical Report 7/1996. The Norwegian Water Resources and Energy Directorate (NVE), ISBN 82-410-0273-4, https://publikasjoner.nve.no/publication/1996/publication1996_07.pdf, 1996.

Saloranta, T. M.: Operational snow mapping with simplified data assimilation using the seNorge snow model, Journal of Hydrology, 538, 314–325, https://doi.org/10.1016/j.jhydrol.2016.03.061, 2016.

Seibert, J. and Bergström, S.: A retrospective on hydrological catchment modelling based on half a century with the HBV model, Hydrology and Earth System Sciences, 26, 1371–1388, https://doi.org/10.5194/hess-26-1371-2022, 2022.

Skaugen, T. and Onof, C.: A rainfall-runoff model parameterized from GIS and runoff data, Hydrological Processes, 28, 4529–4542, https://doi.org/10.1002/hyp.9968, 2014.

Tveito, O.: Norwegian standard climate normals 1991–2020 – the methodological approach, Tech. rep., MET report 5/2021. The Norwegian Meteorological Institute, ISSN 2387-4201, 2021.

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, Hydrology and Earth System Sciences, 19, 913–931, https://doi.org/10.5194/hess-19-913-2015, 2015.

https://doi.org/10.5194/egusphere-2026-1056
Preprint. Discussion started: 6 March 2026

Vormoor, K., Lawrence, D., Schlichting, L., Wilson, D., and Wong, W. K.: Evidence for changes in the magnitude and frequency of observed
555 rainfall vs. snowmelt driven floods in Norway, Journal of Hydrology, 538, 33–48, https://doi.org/10.1016/j.jhydrol.2016.03.066, 2016.

Winsvold, S. H., Andreassen, L. M., and Kienholz, C.: Glacier area and length changes in Norway from repeat inventories, The Cryosphere, 8, 1885–1903, https://doi.org/10.5194/tc-8-1885-2014, 2014.