https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

Evaluating Flexible Configurations of the Shyft Hydrologic Model Framework Across Mainland Norway

Olga Silantyeva[1], Shaochun Huang[2], and Chong-Yu Xu[1]

[1]Department of Geosciences, University of Oslo, Sem Sælands vei 1, Blindern, 0371 Oslo, Norway
[2]Norwegian Water Resources and Energy Directorate (NVE), Middelthuns gate 29, 0368 Oslo, Norway

Correspondence: Olga Silantyeva (olga.silantyeva@geo.uio.no)

Abstract. The development and application of numerous hydrological models have played an indispensable role in advancing our understanding of hydrological processes, improving forecasting capabilities, supporting the design and operation of water conservancy projects, and facilitating water resource assessments. However, due to the spatial heterogeneity and temporal variability of climate and basin characteristics, the inherent complexity of hydrological processes, and data limitations, hydrological modeling faces two major bottlenecks: first, no single model is universally applicable to all river basins; second, further improvement in simulation accuracy of existing fixed-structure models remain challenging. As a result, the emergence of hydrological modeling frameworks with flexible structures and configurable components represents the next generation in the model development. Shyft is one of such flexible modeling frameworks fulfilling the above-mentioned purpose. It is cross platform and open source, jointly developed by academic and industrial partners. The framework allows uncertainty analysis, streamflow simulations, and forecasting. Most evaluation efforts of the framework to date have focused on smaller basins, but there is also a need to benchmark model performance more comprehensively. Here, we present a public benchmark for discharge simulation for 109 catchments across mainland Norway. Five model configurations are evaluated containing two different evapotranspiration routines (Priestley-Taylor and Penman-Monteith), two runoff methods (Kirchner and HBV) and two snow modules (temperature-index and semi-physical). The models are calibrated with 10 variants of target goal functions: KGE-based family, consisting of KGE, LKGE, bcKGE, KGE_LKGE, KGE_bcKGE, and NSE-based family, with NSE, LNSE, bcNSE, NSE_LNSE, NSE_bcNSE. The simulations are divided into two major groups: without precipitation correction and with precipitation correction. The evaluation is performed from 1981 to 2020 (approx.40 years) at a daily time step. Using KGE, NSE and percent bias (PBIAS) as main evaluation metrics, the model configurations are compared against each other and against climatological benchmarks. The results show that all selected models were able to beat both mean and median flow benchmarks for the majority of catchments in all the target goal function set ups. 89% of catchments gain higher performance scores with precipitation correction, but the picture is mixed for different metrics and models. The KGE and NSE performance metrics reveal that models, which combine temperature-index snow-tiles model and Kirchner runoff (**-STK**), perform best, but require precipitation correction to improve PBIAS. The models, which have semi-physical gamma-snow routine (**-GSK**), show relatively low performance with KGE and NSE scores, especially in Mountain and Inland hydrological regimes, but have the lowest |PBIAS| if no precipitation correction is applied. Precipitation correction shows limited effect on the **-GSK** models, even deteriorating some of the scores. The model, which combines temperature index snow-tiles and HBV runoff instead of

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

Kirchner (**-STHBV**), is the most sensitive to precipitation correction: it has the worst PBIAS score across all models without precipitation correction, but jumps to third place in all three metrics, if the correction is applied. The study highlights that KGE-based goal functions reduce PBIAS more than any of the NSE-based goal functions. The study confirms that logarithmic transformation on streamflows, both if LKGE and LNSE are used as target goal functions, generate parameter sets with majority of outliers (KGE scores lower than -0.41). This new benchmark has potential to help with diagnosing problems, improving algorithms and further development within hydrological part of Shyft. Modeling results are made publicly available for further investigation.

1 Introduction

Hydropower has a significant role in the Norwegian energy system, contributing about 90% of the electricity, (www.ssb.no, 2025). Studies show that there is high confidence that the magnitude and seasonality of peak flows in Norway will change due to climate change (Hanssen-Bauer et al., 2017; Nilsen et al., 2022). This will further impact electricity supply and flood risks throughout the regions of Norway. In this context, hydrologic modelling continues to be a fundamental and critically important tool for water resource managers and hydropower operators, as accurate forecasts of inflow help to estimate available hydropower, maintain electricity prices at affordable levels and mitigate flood risks efficiently. However, hydrological modeling is subject to various sources of uncertainty related to input data, calibration, model structure (Moges et al., 2021), or sampling (Knoben et al., 2025). In fact, quantifying uncertainty in hydrological modelling is named as one of the unsolved problems in hydrology (Blöschl et al., 2019). Acknowledging input data uncertainty is an important part of hydrologic analysis and water management (McMillan et al., 2018). Some studies focus on precipitation uncertainty (Bárdossy et al., 2022), others add temperature uncertainty to the analysis (Engeland et al., 2016; Tang et al., 2025). The selection of an objective function for calibration contributes to parameter uncertainty (Onyutha, 2024). Multi-objective analysis is one of the possible approaches to address calibration uncertainty (Moges et al., 2021). Flexibility of model structures in many available frameworks, like the Structure for Unifying Multiple Modeling Alternatives (SUMMA) (Clark et al., 2015) or Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) (Knoben et al., 2019a) allows for comprehensive analysis of model structural uncertainty. This type of uncertainty has recently received more attention, as the type of uncertainty underrepresented in the curriculum of future hydrologists (Knoben and Spieler, 2022).

Shyft is an open-source and fully FAIR (findable, accessible, interoperable, reusable software) framework for uncertainty analysis and hydrologic modelling developed by the Norwegian hydropower company, Statkraft AS, in cooperation with the University of Oslo (Burkhart et al., 2021). The goal of Shyft is to facilitate collaboration among system providers, users, and research communities solving energy market-related problems. The Shyft framework development was largely influenced by the SUMMA approach (Clark et al., 2015), offering a choice of conceptual models for different purposes, while meeting oper-

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

ational requirements for security, efficiency and resilience. Shyft contains different model structures and supports optimizing using multiple goal functions, making it an appealing candidate for a comprehensive hydrological analysis. Even though Shyft is used operationally and in a number of small-scale research studies (Westergren, 2016; Matt et al., 2018; Teweldebrhan et al., 2018; Bhattarai et al., 2020a; Skavang, 2023), a public benchmark showcasing its functionality in a large sample of catchments has so far been lacking. This makes it difficult to compare existing components with newly developed ones, identify their strengths and weaknesses, and progress with new development within the framework.

Benchmarking of a hydrological model is an exercise to assess the applicability of the model for various purposes and is an emerging trend in the hydrological modelling community (Beven, 2023; Newman et al., 2017; Knoben et al., 2020; Towler et al., 2023). Gupta et al. (2014) proposed large-sample hydrological studies as an approach for understanding catchment processes with modelling at a variety of hydrological regimes, spatiotemporal scales and environments. This work has been further supported with the development of datasets like CAMELS for the US (Addor et al., 2017) and more recently the global Caravan dataset (Kratzert et al., 2023). Recent studies demonstrated the potential of large-sample hydrology in Norway – for analysing streamflow sensitivity to air temperature (Hegdahl et al., 2019), understanding droughts (Bakke et al., 2020), estimating potential evaporation and evaluating model performance (Huang et al., 2019), exploring regional trends and extremes (Yang and Huang, 2023).

Using Shyft as an example, the objective of this research is to evaluate the performance of flexible model configurations from a benchmarking perspective, considering different objective functions, accuracy of precipitation input, and streamflow regimes. We evaluate Shyft for its ability to predict streamflow for a large set of catchments of different sizes across mainland Norway with a variety of hydroclimatic regimes. We are interested in understanding model limitations and opportunities for further development and providing a publicly available benchmark. In addition, we discuss how to define the proper strategies for choosing the goal function and model structure and their limitations for large sample hydrological analysis, so the model set-up is fit-for-purpose.

The rest of the paper is structured as follows: section 2 describes the study area and forcing data. Section 3 presents the hydrological model, performance metrics and experimental design. Section 4 demonstrates the simulation results. Section 5 provides discussion of the results, shortcomings and future work. Conclusion is presented in section 6.

2 Study area and Data

2.1 Study area

Mainland Norway is a country in Northern Europe, spanning latitudes from 58^o to 71^o North and covering 324220 km^2. Norway exhibits significant variation in both topography and climate. Beck et al. (2018) provide a detailed climate classification: the west coast has a temperate oceanic climate with high precipitation rates; further east, the climate shifts to the cold type with no dry season and cold summer, characterised by lower precipitation and greater seasonal temperature variation. High mountain areas are classified as Polar tundra, while the south coast has a cold type of climate with no dry season and warm summers (see Fig. 1).

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

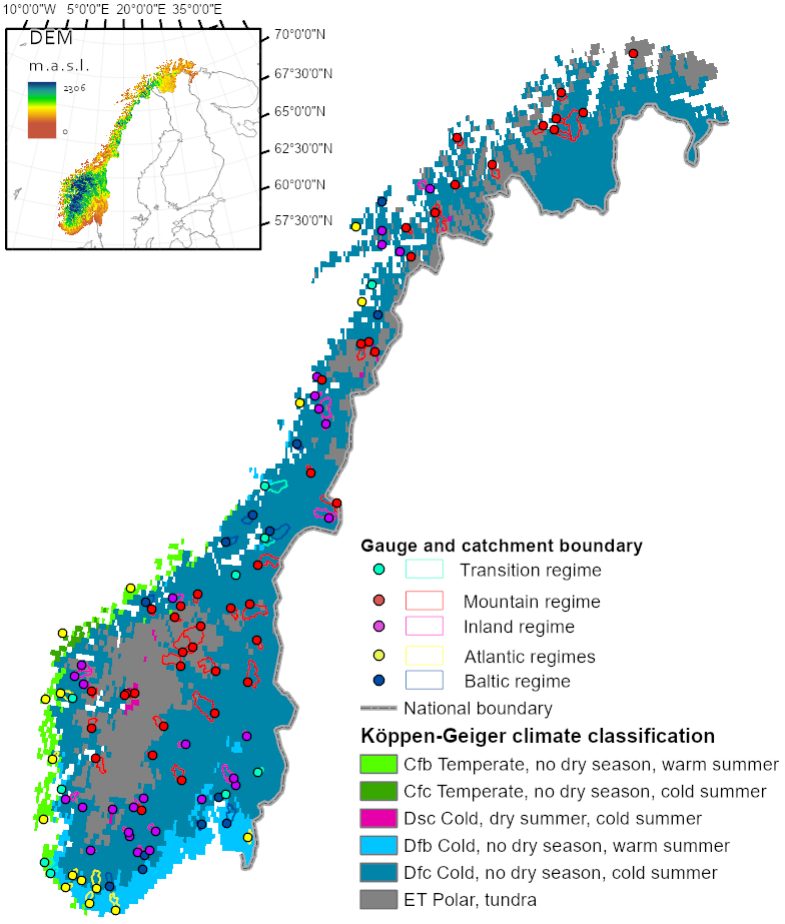

Figure 1. Study area: The location and hydrological regimes of the 109 catchments based on the definitions in (Bakke et al., 2020) and climate regimes in Norway according to Köppen-Geiger climate classification (Beck et al., 2018).

The study area contains 109 catchments from mainland Norway (see Fig. 1). The catchments were selected based on the following criteria: no regulation, less than 5% missing data in observed discharge for the selected study period (from 1981 to 2020), excluding catchments larger than 1000 km^2. The catchments are spread through the country, capturing the full range of Norway's climatic and hydrological regimes.

To define the runoff regimes for the selected catchments we refer to the study by Bakke et al. (2020). We identify the following characteristic runoff regimes:

- (a) Mountain regime – characterized by low flow during two months in winter or early spring due to snow accumulation, followed by spring or early summer high flow driven by snowmelt.
- (b) Inland regime – similar to the Mountain, but with an additional runoff peak in autumn caused by rainfall.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

- (c) Atlantic regime – characterized by high runoff during autumn or winter caused by rainfall, and low flow during spring and summer driven by high evapotranspiration, low precipitation, or both.
- (d) Baltic regime – similar to the Atlantic regime, but has additional high flow period during early spring related to snowmelt.
- (e) Transient regime – intermediate type, representing mixture of Inland and Baltic characteristics.

The regimes are shown in Fig. 1. The same definition of regimes is used, for example, in Yang and Huang (2023). For the majority of catchments in this study the hydrological regime is classified as the Mountain (43 out of 109). The regime in 27 catchments is classified as Inland, and 16 catchments exhibit characterisitcs of the Atlantic regime. Only 11 and 12 catchments fall into the Baltic and Transient categories, respectively. As the Mountain regime is more representative, we should interpret the following results for other regimes cautiously, noting the limited sample size of other regimes.

2.2 Data

The forcing dataset is seNorge2018 (Lussana et al., 2019) for precipitaiton and mean temperature and hySN5 (Erlandsen et al., 2019) for relative humidity and radiation at daily time resolution. Wind data comes from the NOrwegian ReAnalysis 10 km (NORA10) product (Reistad et al., 2011). The data is at 1km spatial scale and daily timestep. Observed discharge at daily timestep for stations is provided by Norwegian Water and Energy Directorate (NVE) and is freely available at https://seriekart.nve.no.

Shyft uses high resolution triangular-irregular network (TIN) mesh, generated from 10 m digital elevation map (DEM) from the Norwegian Mapping Authority (https://hoydedata.no/LaserInnsyn2/) with 80 m Corine land cover map (https://land.copernicus.eu/global/products/).

3 Methods

3.1 Shyft hydrological model

Shyft is a cross-platform, open source toolbox (https://gitlab.com/shyft-os/shyft). Shytf.hydrology is a component of the Shyft framework focused on hydrological modelling. Description of the previous Shyft version is provided in (Burkhart et al., 2021). Figure 2 shows plug and play components within the system. One can combine the components into full conceptual model, which in the Shyft ecosystem is called **stack**. Typical stack takes as input forcings: precipitation, temperature, relative humidity, wind speed and shortwave radiation, interpolates them into the gridded cells using, for example, inverse distance weighting (IDW). The short-wave incoming radiation can be also adjusted according to slope and aspect information from the cell. The processed forcings are further combined into *region environment* and attached to *region model*, which has components defining evapotranspiration, snow modelling, glacier melt, streamflow model and river routing. The *region model* contains simulation

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

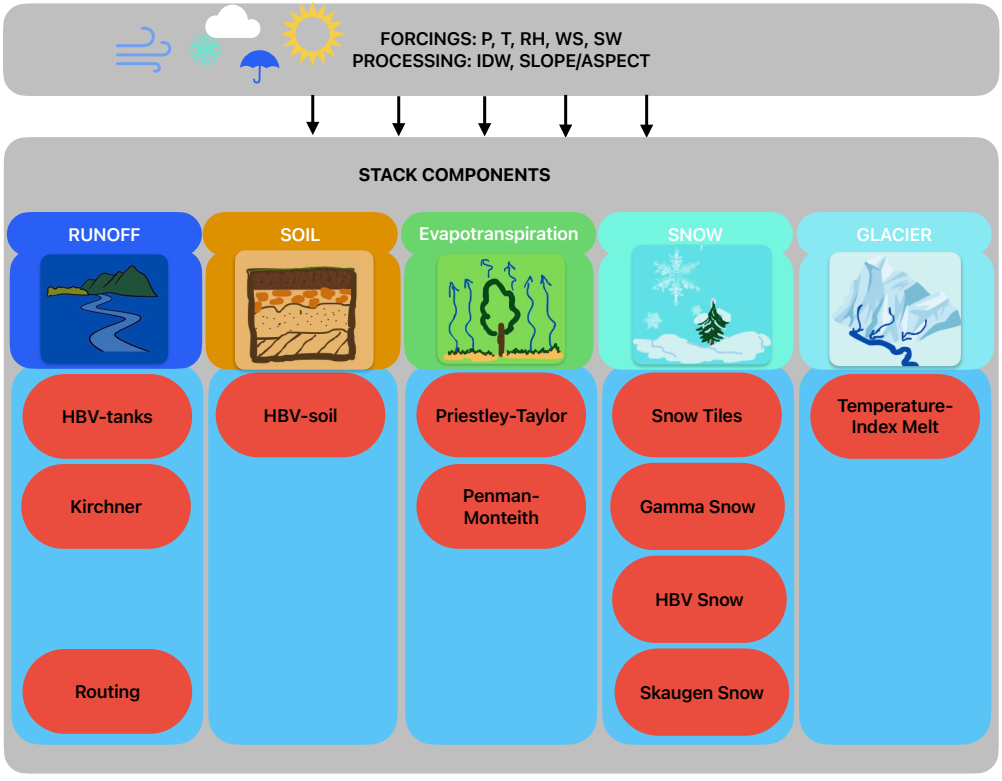

Figure 2. Shyft stack components in Shyft.hydrology: The model interpolates forcings (temperature, precipitation, shortwave radiation, relative humidity and wind speed) with a selection of algorithms. The stack is a conceptual hydrological model, which defines: evapotranspiration, snow response, glacier melt, soil moisture and runoff response and routing. For further details, please refer to (Burkhart et al., 2021)

domain, which might be represented as lumped, square cells or TINs. Though, Shyft allows data assimilation, this is not covered in this paper.

We test here 5 model stacks (see Table A1): **PTSTK**, used in Skavang (2023); completely new **RPMSTK**; **PTGSK**, which is previously defined as best performing (Teweldebrhan et al., 2018); **RPMGSK**, used in Bhattarai et al. (2020b); and new **PTSTHBV**. The **PTSTK** is the simplest model in the selection with **P**riestley-**T**aylor routine for evapotranspiration modelling (Priestley and Taylor, 1972), **S**now-**T**iles temperature index snow model, described in Skavang (2023) and **K**irchner routine for streamflow (Kirchner, 2009). The next level of complexity comes with **R**adiation correction on the slopes algorithm (Allen et al., 2006) and **P**enman-**M**onteith evapotranspiration modeling (Dingman, 2015) in **RPMSTK** stack. The snow modeling in shyft has a simplified energy-balance **G**amma-**S**now algorithm, which is a component of **PTGSK** and further modification **RPMGSK** stack. We also test **PTSTHBV** stack, which uses HBV-soil and HBV-tank (Bergstrom, 1991; Bergström and Lindström, 2015) instead of Kirchner runoff method.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

In this study the cells are represented by TIN mesh created using open source software rasputin v.0.3.alpha (https://github.com/expertanalytics/rasputin/), which is described in (Silantyeva et al., 2023). An example of TIN-mesh for one of the catchments in this study is shown on the Fig. A1. It has app. 0.1 km^2 average cell size.

Table A1 shows model configurations used for the study and parameters involved, where "x" – indicate that parameter exists in the model. River routing was off for all of the models. Glacier melt uses simple temperature index model and is kept at default values for all configurations. Explanation of parameters can be found in Skavang (2023); Silantyeva et al. (2023) and Lawrence et al. (2009) for HBV related part. The parameters indicated with minimum and maximum values – are the calibrated parameters, the rest are not calibrated.

3.2 Performance metrics

Studies suggest multi-criteria model evaluation (Cinkus et al., 2023; Onyutha, 2024), thus, we use several criteria for evaluation process: The Kling–Gupta efficiency (KGE) is used as the overall performance metric, accompanied by Nash-Sutcliffe Efficiency (NSE), which remains here, for easier comparisons with other studies and also as a default performance metric within operations of the hydropower companies. Percent bias (PBIAS) is added to the two metrics to assess, if the models tend to over (positive sign) or under (negative sign) estimate streamflow volumes.

The KGE is defined as follows (Gupta et al., 2009):

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\alpha-1)^2},$$

where r is the linear Pearson correlation coefficient between simulation (sim) and observation (obs), $\beta = \frac{\theta_{sim}}{\theta_{obs}}$ is ratio of standard deviation of streamflow θ, $\alpha = \frac{\mu_{sim}}{\mu_{obs}}$ is the ratio between means of the flow μ. The three components of KGE reflect the similarity between simulations and observations in terms of the correlation between the two flows: r-term, bias: β-term, and variability α-term (Gupta et al., 2009).

The NSE is defined as follows (Nash and Sutcliffe, 1970):

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{N} \left(Q_{obs}^{t} - Q_{sim}^{t}\right)^2}{\sum_{t=1}^{N} \left(Q_{obs}^{t} - \overline{Q_{obs}}\right)^2},$$

where Q_{obs}^{t} is observed discharge and Q_{sim}^{t} is simulated discharge at time t, $\overline{Q_{obs}}$ is mean of observed discharge.

KGE and NSE range between -∞ to 1, where 1 indicates a perfect match. The usability of both metrics is under an intense discussion in the hydrologic community with KGE, being the most recommended metric, as it takes into account flow variability (Pushpalatha et al., 2012; Knoben et al., 2019b, 2020; Althoff and Rodriguesa, 2021; Yang et al., 2022).

The PBIAS metric is defined as:

$$\mathrm{PBIAS} = 100 \cdot \frac{\sum_{t=1}^{N} \left(Q_{sim}^{t} - Q_{obs}^{t}\right)}{\sum_{t=1}^{N} Q_{obs}^{t}},$$

In the results section we use |PBIAS|, which is expected to be within 15% range, (Moriasi et al., 2007). In addition, we use med(KGE), med(NSE), med(PBIAS) as a notation for median of the set.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

We also use a combined criteria for model evaluation, indicating runs where KGE>-0.41 (Knoben et al., 2019b), NSE>0.0 (Knoben et al., 2019b) and |PBIAS|<15% (Moriasi et al., 2007) are satisfied together.

3.3 Goal functions

Goal functions play an important role in the calibration of the hydrological model. The goal functions in Shyft are limited to KGE, NSE and RMSE, their transformations and combinations. We defined KGE-based goal functions group, consisting of KGE, LKGE (KGE calculated on log-transformed discharge), bcKGE (KGE calculated with box-cox transformation of flow, where λ=0.3, (Santos et al., 2018)), KGE_LKGE $= \frac{\text{KGE+LKGE}}{2}$, KGE_bcKGE $= \frac{\text{KGE+bcKGE}}{2}$ and NSE-based group, with NSE, LNSE (calculated on log-transformed discharge), bcNSE (NSE calculated with box-cox transformation of flow, where λ=0.3), NSE_LNSE $= \frac{\text{NSE+LNSE}}{2}$, NSE_bcNSE $= \frac{\text{NSE+bcNSE}}{2}$.

The recent study by Thirel et al. (2023) shows that transformations such as logarithmic and box-cox can give a valuable generalist performance metric, showing good results for the intermediate range of flows and acceptable results for high and low flows. However, there are hidden numerical problems with LKGE (Santos et al., 2018), which might impact convergence of the optimisation algorithm and show misleading low performance result.

Thus, we run calibration towards 10 goal functions: KGE, LKGE, bcKGE, KGE_LKGE, KGE_bcKGE, NSE, LNSE, bcKGE, NSE_LNSE and NSE_bcNSE with local optimisation algorithm: bobyqa (Powell, 2009). The table A1 shows parameters of each model configuration.

3.4 Experimental design

We run experiments for a large sample of catchments and a set of stacks: **PTSTK** and **RPMSTK**, **PTGSK** and **RPMGSK**, and **PTSTHBV**. We use bold uppercase notation for the model names in the text, but in the figures the notation is changed to small case. We set climatological benchmark as interannual mean and median flow per calendar day for each catchment, as described in Knoben et al. (2020) and applied, for example, in Towler et al. (2023). The **PTSTK** is a modeling benchmark, as it represents the simplest conceptual model available in Shyft. We adopt split-sample calibration (years 1981-2000) and validation (years 2001-2020) approach using one year of warm-up period prior to calibration. In the first set of experiments, the precipitation correction is not allowed. Thereas, last set of experiments is run with precipitation correction. This can help us assess the influence of precipitation accuracy on the model performance and model robustness, and get insights on the uncertainty of the input data.

4 Results

All simulations are available in the Zenodo archive: Silantyeva and Huang (2025). Here we group the simulations based on the stacks, goal functions, with/without precipitation corrections and hydrological regimes to analyse the model performance for each group. The boxplots, presented here, do not show all outliers, excluding those, from non-converged runs (NSE or KGE scores $\rightarrow \infty$).

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

4.1 Sensitivity to target goal function selection

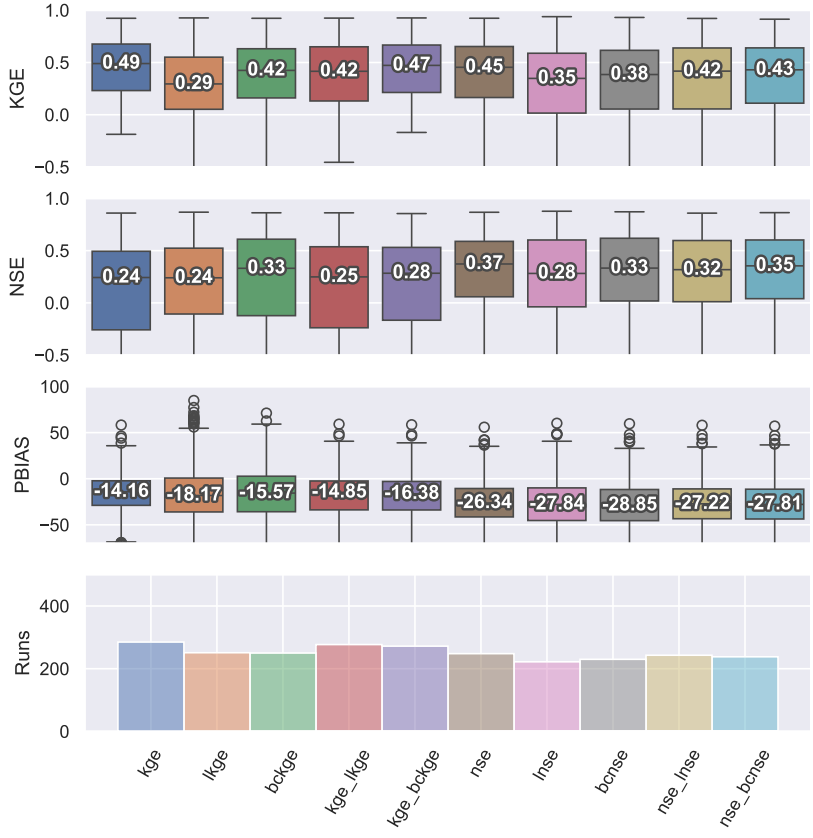

Figure 3. Performance of goal functions **without precipitation corrections** averaged for all catchments and all stacks. Validation and calibration periods are considered separate runs. From top to bottom: KGE, NSE and PBIAS scores for each model; Last row: number of runs, where the model satisfies three criteria: KGE>-0.41, NSE>0.0 and |PBIAS|<15%.

Figure 3 presents box plots summarizing the distribution of KGE, NSE and PBIAS scores across the goal functions averaged for five stacks. In addition, histogram containing number of runs, satisfying three criteria simultaneously: KGE>-0.41, NSE>0.0 and |PBIAS|<15% is shown. For each goal function, the results from all model runs (calibration and validation considered separate runs) without precipitation correction are combined together, giving 10 runs for each goal function for each catchment, which gives 1090 runs per goal function. As can be noticed, there are best performing goal function options in each of the performance metric. For KGE performance metric, KGE and KGE_bcKGE produce highest med(KGE) of 0.49 and

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

0.47, followed by NSE (med(KGE) = 0.45) and NSE_bcNSE achieving med(KGE)=0.43. LKGE goal function gives lowest med(KGE)=0.29 across all goal functions. LNSE is the next poor performing option with med(KGE)=0.35.

For the NSE performance metric, NSE as a goal function produces highest med(NSE)=0.37 followed by NSE_bcNSE with med(NSE)=0.35. LNSE has lowest median value (med(NSE)=0.28) and highest spread between NSE-based goal functions. The best performing goal function in the KGE-based family is bcKGE (med(NSE)=0.33), followed by KGE_bcKGE (med(NSE)=0.28). All KGE-based goal functions demonstrate relatively wider spread in the values compared to NSE-based goal functions.

The correlation coefficient between PBIAS and KGE is >0.99. Low KGE or even NSE is caused by the large bias. As can be seen, for the PBIAS performance metric, all KGE-based goal functions give highest med(PBIAS) values with KGE and KGE_LKGE slightly above the rest of the family. The NSE-based goal functions show similar med(PBIAS) values, lower, compared to KGE-based goal functions. The large negative bias may result from the bias of precipitation forcing.

When it comes to satisfying three criteria (KGE>-0.41, NSE>0.0, |PBIAS|<15%) the ranking for the best goal functions is: KGE, KGE_LKGE, KGE_bcKGE. LNSE as a goal function generates lowest number of runs satisfying three criteria.

4.2 Sensitivity to model configurations

Figure 4 summarizes KGE, NSE and PBIAS scores for each of the selected models. For this analysis for each available model we combine all goal functions together, having a total population of 2180 runs per model.

As can be seen, **RPMSTK** model is best performing model with KGE and NSE meric (med(KGE) = 0.68, med(NSE)=0.64), followed by **PTSTK** (med(KGE) = 0.59, med(NSE)=0.56). Interquartile ranges (IQR) for these two models are lowest, but the whiskers indicate that there are outliers in the population. The **RPMGSK** model has med(KGE)=0.27, slightly above **PTGSK** (med(KGE)=0.21) and **PTSTHBV** (med(KGE)=0.15). The IQR for **PTGSK** and **RPMGSK** models suggest high spread in the data, with significant amount of outliers, but for **PTSTHBV** the data is very narrowly distributed around the median value. With NSE metric **PTSTHBV** is ranked three between models with med(NSE)=0.22, whereas both **RPMGSK** and **PTGSK** models got negative median scores with notable spread. All models significantly underestimate the streamflow. The **RPMGSK** model has the best results of med(PBIAS)=-7.67%, followed by **RPMSTK** (med(PBIAS)=-15.25). The **PTSTHBV** model has the worst results between models: med(PBIAS)=-50.1%.

Last row of the Fig. 4 can be used to perform final ranking of the models. The **RPMSTK** model has highest number of runs satisfying the three criteria, followed by **PTSTK**. The **RPMGSK** model is slightly behind **PTSTK** and above **PTGSK**. The **PTSTHBV** model has lowest number of runs satisfying all three criteria. This final ranking is same as if the only KGE was used as performance score. However, less than half of the runs satisfy all three criteria simultaneously.

Figure 5 demonstrates cumulative density functions (CDFs) for KGE scores for the variant without precipitation correction. We separate calibration and validation runs in the figure to demonstrate temporal transferability of the models. Seasonal benchmark for median discharge (med(Q)) is shown in pink and for mean discharge (mean(Q)) in orange. Both mean(Q) and med(Q) as benchmarks demonstrate rather low med(KGE) scores of -0.01 and 0.16 respectively. More than 43% of catchments have

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

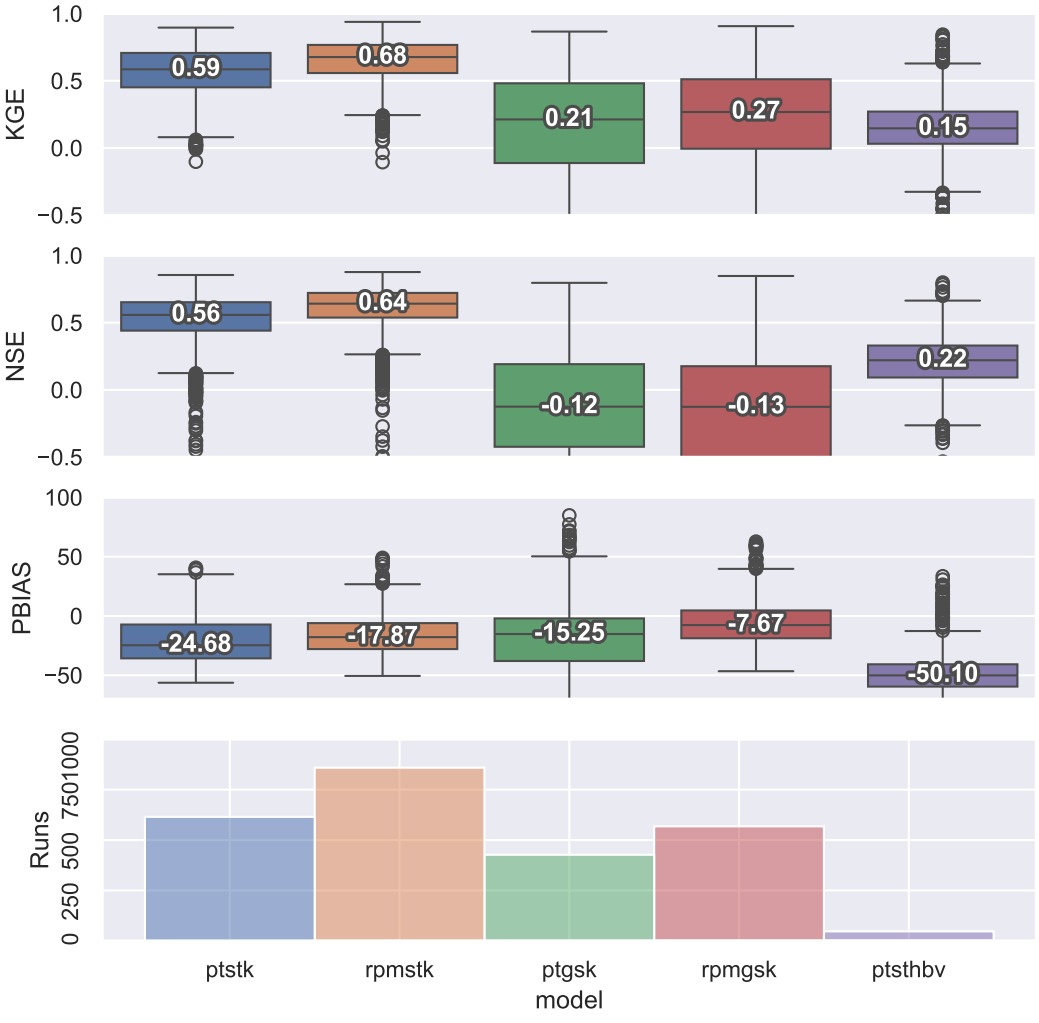

Figure 4. Performance of models for runs **without precipitation corrections** averaged for all catchments and 10 goal functions. Validation and calibration periods are considered separate runs. From top to bottom: KGE, NSE and PBIAS scores for each model; Last row: number of runs , where model satisfies criteria: KGE>-0.41, NSE>0.0 and |PBIAS|<15%.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

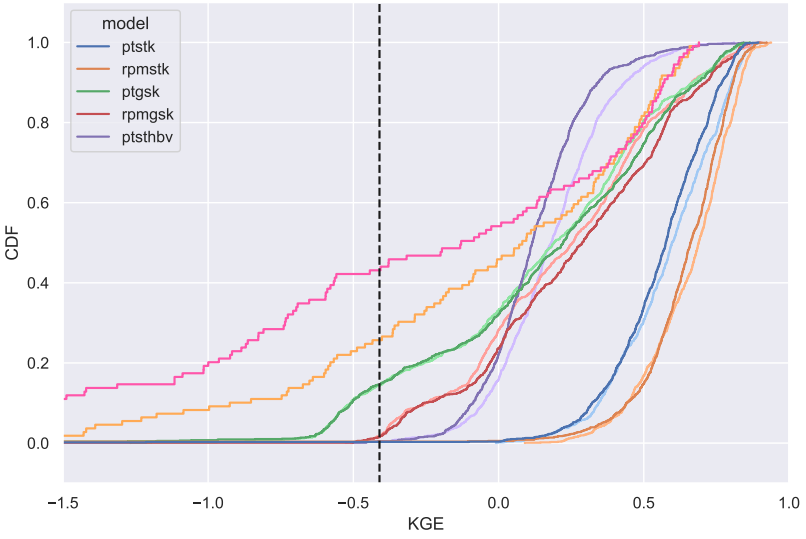

Figure 5. Cumulative density functions (CDFs) for Kling–Gupta efficiency (KGE) scores based on daily streamflow without precipitation correction. The blue lines correspond to **PTSTK** model, the orange lines correspond to **RPMSTK**, the green lines – to **PTGSK**, the red lines – **RPMGSK** and purple lines – **PTSTHBV**, where the calibration period has lighter color. The dotted vertical line is the KGE mean flow benchmark (-0.41) (Knoben et al., 2019b). Seasonal benchmark for median Q is shown in pink and for mean Q in orange.

KGE score lower than -0.41 when mean discharge is used as a flow predictor. The median discharge perform slightly better with less than 26% of catchments lying lower than -0.41 benchmark.

Same as the previous result based on combined criteria, the comparison with seasonal benchmarks confirms that **RPMSTK** is the best performing model, followed by **PTSTK**. KGE scores for **RPMGSK** and **PTGSK** are behind the top performing group, but still higher than the seasonal benchmarks scores. The **PTSTHBV** model is the only one, which has scores lower than the seasonal benchmarks, intersecting both med(Q) and mean(Q) benchmarks. All models demonstrate strong temporal transferability, performing in validation period similar to calibration. However, the **PTGSK** and **RPMGSK** are models, where certain catchments showed better performance in validation period.

4.3 Sensitivity to precipitation correction

SeNorge2018 is a gridded 1x1km precipitation and temperature dataset based on observations from surface meteorological stations (Lussana et al., 2019). Some regions, especially mountainous, have limited number of precipitation gauges, which contributes to larger bias. In this case, a correction is usually necessary. To this point we showed results of simulations, where precipitation correction was not allowed during calibration. This section presents results of the simulations with precipitation correction. The precipitation correction factor was allowed during calibration in the range [0.4;2.0] based on our expert

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

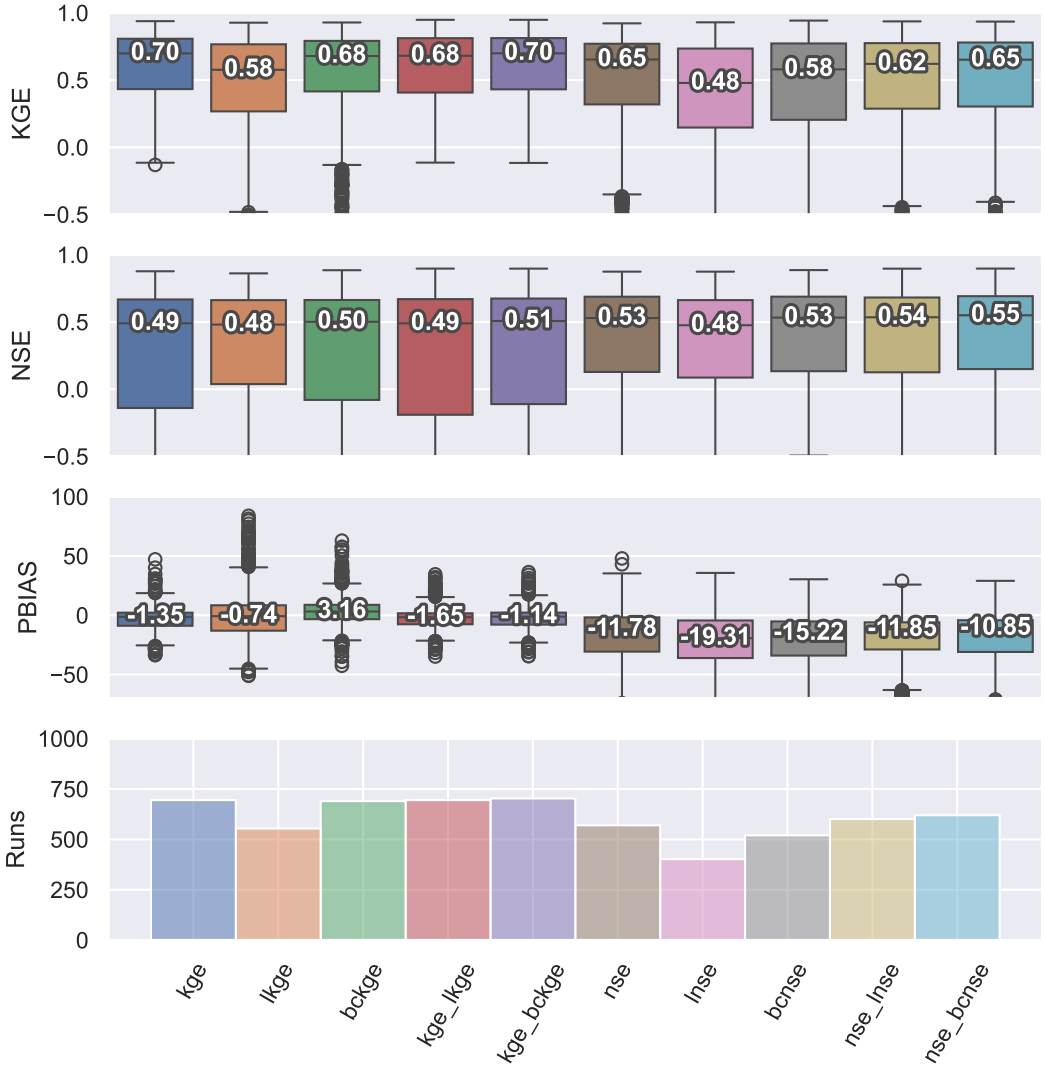

Figure 6. Performance of goal functions **with precipitation corrections** averaged for all catchments and all stacks. Validation and calibration periods are considered separate. From top to bottom: KGE, NSE and PBIAS scores for each model; Last row: number of runs, where the model satisfies three criteria: KGE>-0.41, NSE>0.0 and |PBIAS|<15%.

255 knowledge. The precipitation correction adjusts precipitation amount based on the simulated discharge as a simple scaling factor.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

Figure 6 demonstrates perfromance scores (KGE, NSE and PBIAS) for each goal function setting and in the last row: histogram with number of runs for each setting, where the model satisfies combined criteria (KGE>-0.41, NSE>0.0, |PBIAS|<15%) for the case with precipitation correction.

The median KGE scores improved for all goal functions with KGE and KGE_bcKGE getting highest values (med(KGE)=0.70), followed by KGE_LKGE (med(KGE)=0.68). The LNSE generates lowest med(KGE)=0.48 and highest spread among all goal functions, outperformed by bcNSE, LKGE and bcKGE with med(KGE)=0.58.

The NSE metrics are also improved with precipitation correction among all goal functions. The NSE_bcNSE produces runs with highest med(NSE)=0.55, followed by NSE_LNSE (med(NSE)=0.54), followed by NSE and bcNSE (med(NSE)=0.53). The LNSE, LKGE and bcKGE as goal functions generate runs with lowest med(NSE)=0.48, slightly outperformed by KGE and KGE_LKGE (med(NSE)=0.49), but the latter two have higher spread.

The PBIAS metric showed a prominent change with precipitation correction. All KGE-based goal functions got median values between -0.74 and 3.16, with significantly reduced IQR. The NSE-based goal functions also improved PBIAS score with precipitation correction, but still significantly underestimated streamflow with med(PBIAS) between -10.85% (NSE_bcNSE) to -19.31% (LNSE).

The last row of the Fig. 6 has now higher number of runs satisfying three criteria compared to the case without precipitation correction. The KGE, KGE_bcKGE as goal functions generate highest number of runs satisfying criteria, just slightly outperforming KGE_LKGE and bcKGE. Same as in the case without precipitation correction, LNSE as a goal function has lowest number of runs within criteria. bcNSE is slightly above LNSE and behind LKGE, which has lowest number of runs within criteria for KGE-based goal functions. Overall, as we can see KGE-based goal functions are more likely to produce results with KGE>-0.41, NSE>0.0 and |PBIAS|<15% simultaneously.

Figure 7 summarizes KGE, NSE and PBIAS scores for each model configuration. Compared with the results from the previous section without precipitation correction, **RPMSTK** still shows the best performance among the five models (med(KGE)=0.81, med(NSE)=0.70 and med(PBIAS)=-2.11%), followed by **PTSTK** (med(KGE)=0.78, med(NSE)=0.65, med(PBIAS)=-2.75%). The IQR for the models narrowed around the median value with limited number of outliers. The performance of **PTSTHBV** has been significantly improved using precipitation correction, with med(KGE)=0.62, med(NSE)=0.55 and med(PBIAS)=-5.92%, it is number three in each of the separate metrics. The **PTGSK** and **RPMGSK** models does not show any improvements in the scores. Furthermore, with the correction applied, the **RPMGSK** model showed even deteriorated PBIAS score (med(PBIAS)=-13.87).

Last row of the Fig. 7 demonstrates final rank of the models. Compared to the case without precipitation correction, the **RPMSTK** and **PTSTK** models still have highest number of runs satisfying three criteria, but the number three model is now **PTSTHBV**. **RPMGSK** model is slightly behind **PTGSK** and has now lowest number of runs satisfying all three criteria.

Looking into CDF plots on Fig. 8 we can see that all models now predict discharge better than both mean(Q) and med(Q) seasonal benchmarks. The slope of the **PTGSK** and **RPMGSK** models remain similar to the slope of seasonal benchmark, but now the two models are virtually identical in the score. The CDF curve of the **PTSTHBV** model significantly changed the slope, becoming similar to **PTSTK** and **RPMSTK**. The steeper slopes suggest that much of the population got similar performance

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

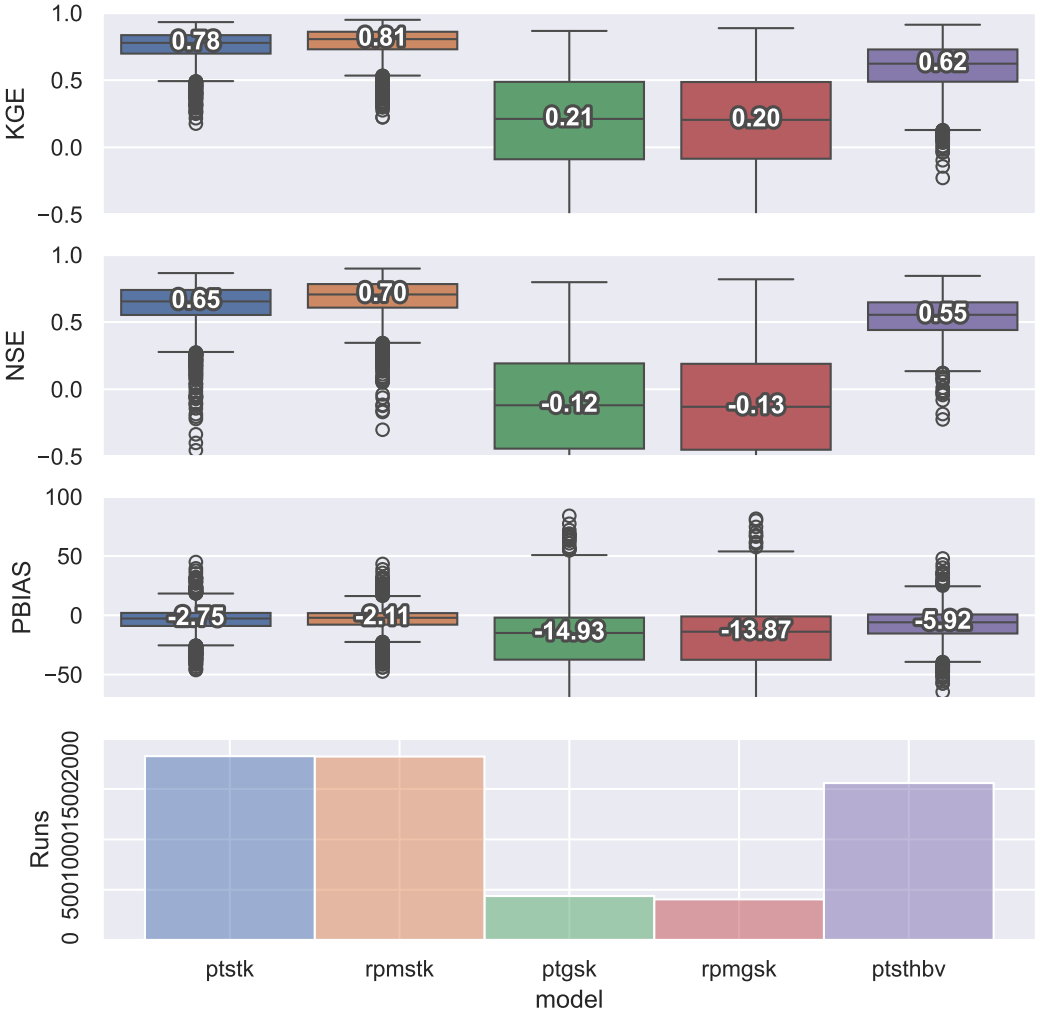

Figure 7. Performance of models for runs **with precipitation corrections** averaged for all catchments and 10 goal functions. Validation and calibration periods are considered separate runs. Left column: KGE, NSE and PBIAS scores for each model; Right column: number of runs, where model is better than benchmark, which is defined as KGE=-0.41, NSE=0.0 and |PBIAS|<15%.

scores. The three models also lack now negative tail, which was visible in the plots without precipitation correction, indicating lack of non-converged calibration runs. Similar to the case without precipitation correction, all models demonstrate strong

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

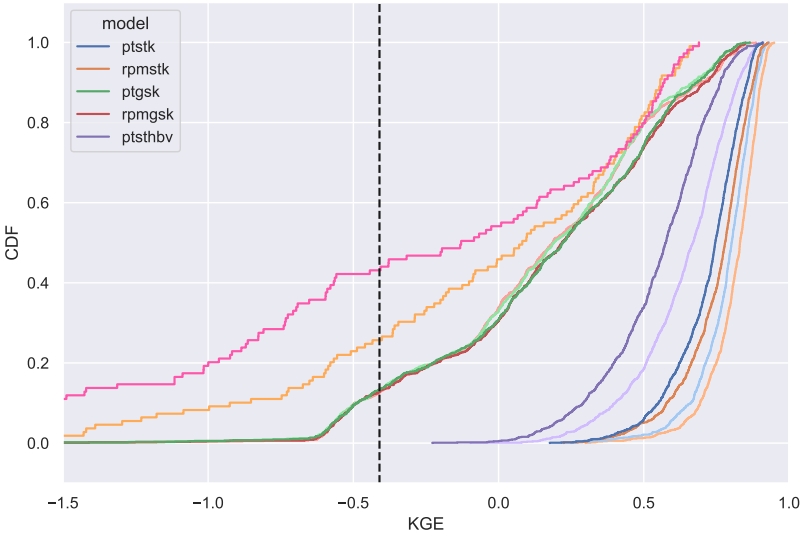

Figure 8. Cumulative density functions (CDFs) for Kling–Gupta efficiency (KGE) scores based on daily streamflow with precipitation correction. The blue lines correspond to **PTSTK** model, the orange lines correspnds to **RPMSTK**, the green lines – to **PTGSK**, the red lines – to **RPMGSK** and purple lines – to **PTSTHBV**, where the calibration period for each model is in lighter color. The dotted vertical line is the KGE mean flow benchmark (-0.41) (Knoben et al., 2019b). Seasonal benchmark for median Q is shown in pink and for mean Q in orange.

temporal transferability, performing in validation period similar to calibration. Again, the **PTGSK** and **RPMGSK** are models, where certain catchments show better performance in validation period.

4.4 Model performance for hydrological regimes

We summarize the model performance results for each of the five regimes using same boxplot structure as before, but there is no combined scores historgram due to differences in catchment populations.

From Fig. 9 we can identify the model ranks for the Mountain regime with and without precipitation. In both cases, the **RPMSTK** has highest KGE and NSE scores, followed by **PTSTK** model. The third place is obtained by **PTSTHBV**, but the latter shows worst PBIAS score without precipitation correction. The three models significantly improve their scores with precipitation correction, but **PTSTHBV** has the most drastic change from med(KGE)=0.09 to 0.67, med(PBIAS) improved from -58.49% to -8.09%. **PTGSK** and **RPMGSK** have lowest KGE ans NSE scores with significant spread with and without precipitation correction, but the PBIAS for **RPMGSK** is best among the models, when precipitation is not corrected. This model shows deteriorated results with precipitation correction. Overall, **RPMSTK** and **PTSTK** are recommended models for the regime, but precipitation should be corrected in order to obtain minimal |PBIAS|.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

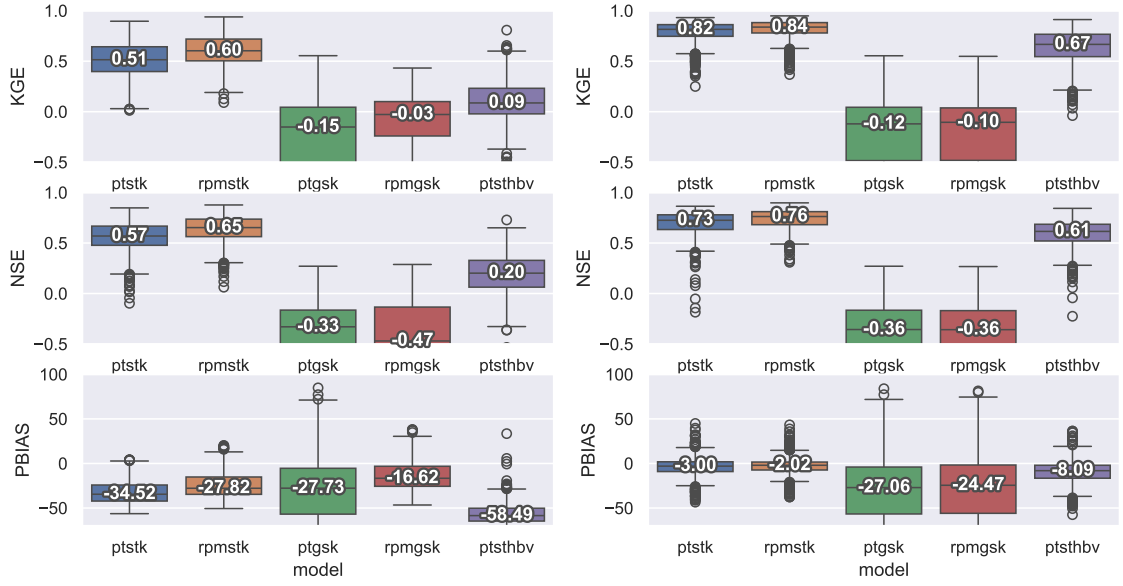

Figure 9. Kling-Gupta efficiency (KGE), Nash-Sutcliffe efficiency (NSE) and Percent Bias (PBIAS) scores based on daily streamflow for catchments classified as Mountain. Performance scores are averaged for all catchments and 10 goal functions. Validation and calibration periods are considered separate runs. Left column: no precipitation correction; right column: with precipitation correction.

For the Inland regime the Fig. A2 demonstrates that the two best performing models remain **RPMSTK** and **PTSTK** in both precipitation cases. Without precipitation correction **RPMGSK** can be considered number three, as it shows med(KGE)=0.29, which is above **PTGSK** and **PTSTHBV**, and has best PBIAS between all the models (med(PBIAS)=-4.62). The **PTSTHBV** model, even thought it shows positive med(NSE)=0.24, significantly underestimate flow with med(PBIAS)=-45.61% . As in the Mountain regime, the precipitation correction improve all three scores for **RPMSTK** and **PTSTK** in the Inland regime with noticable increase in PBIAS. The hypersensitivity of the **PTSTHBV** model to precipitation correction is confirmed for this hydrological regime as well. Again, **RPMGSK** decreases the med(PBIAS) score, when the correction is applied.

The Atlantic regime on the Fig. A3 shows different picture compared to the Mountain and Inland. Without precipitation correction best performing model with KGE and NSE metric is **RPMSTK** (med(KGE)=0.72), but **RPMGSK** and **PTGSK** are just slightly behind (med(KGE)=0.7 for each), though the med(NSE) for this 2 models is lower than for **PTSTK** and **RPMSTK**, the two models outperfom the rest of the group in PBIAS metric. In addition, the results for the models are closely distributed around median. The **PTSTHBV** model remains worst performing in all metrics, including NSE, and its med(PBIAS)=-43.54% way below acceptable range. For this regime precipitation correction leads to improvement of scores for **RPMSTK** and **PTSTK** models, moving the latter to the second place. Though **PTSTHBV** improved its score, its med(KGE)=0.67 still below **PTGSK** (med(KGE)=0.70) and **RPMGSK** (med(KGE)=0.70), but the model again significantly improved its PBIAS score to

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

med(PBIAS)=-0.26, becoming number one with this metric. Thus, for this regime, if precipitation correction is not applied, the **PTGSK** and **RPMGSK** models can be recommended, as they have notably lower PBIAS, than **RPMSTK**. If precipitation is corrected, all models are acceptable, but **RPMSTK** and **PTSTK** are the best.

For the Baltic regime we can see on the Fig. A4: in case there is no precipitation correction, model ranking is similar to that in Atlantic with **RPMSTK** being the best model, acceptable performance of **RPMGSK** and **PTGSK** models and very low med(PBIAS) for **PTSTHBV**. With precipitation correction, the **PTSTHBV** model gets third place, significantly improving all scores.

For the Transient regime the Fig. A5 demonstrates same model rankings as for the Baltic regime, but the noticable difference between regimes, is that even without precipitation correction med(|PBIAS|) for **RPMSTK** and **PTSTK** is relatively low, and this is the only regime, where one of the models (**RPMGSK**) slightly overestimate flow with med(PBIAS)=3.34. For this regime, precipitation correction improves PBIAS metric only for the **PTSTHBV** model, thereas all the rest got even deteriorated results. The **RPMSTK** and **PTSTK** only moderatly improve KGE and NSE scores with precipitation correction. Thus, this is the only regime, where precipitation correction is not recommended for the models.

4.5 Spatial distribution of KGE performance scores

Figure 10 summarizes best KGE results for each catchment without precipitation correction. The **RPMSTK** model is the best for majority of the catchments (85), followed by **PTSTK** (16) spanning all regimes. The **PTSTHBV**, as expected, has not been selected as the best in any of the catchments. **PTGSK** and **RPMGSK** are best performing models for 5 and 4 catchments on the south coast and west coast, respectively. The ranking of the goal functions is: 1. KGE (34), 2. KGE_bcKGE (18), 3. bcKGE (17), 4. NSE (10), 5. LKGE (9), 6. KGE_LKGE (7), 7. LNSE (7), 8. bcNSE (5) and NSE_bcNSE and NSE_LNSE on the 9th and 10th places with 2 and 1 times best goal function choice. The best KGE values span from minimum 0.50 to maximum 0.94, with a mean value of 0.75.

Figure 11 summarizes best KGE results for each catchment with precipitation correction. Only **RPMSTK** (103) and **PTSTK** (7) are selected as the best models, if precipitation is corrected. The ranking of the goal functions changed with top three being: 1. KGE_LKGE (36), 2. bcKGE (20), 3. LKGE (16). NSE as a goal function was not selected as the best for any of the catchments. The best KGE scores in this case span from minimum 0.50 to maximum 0.95, with a mean value of 0.85.

The important takeaway is that there is always a combination of model and goal function, which leads to a good or excellent KGE score. On Fig. A6 and Fig. A7 we show spatial distribution of KGE score groups. As can be noticed without precipitation correction many mountain catchments get low KGE scores, but with precipitation corrected these catchments show improved scores.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

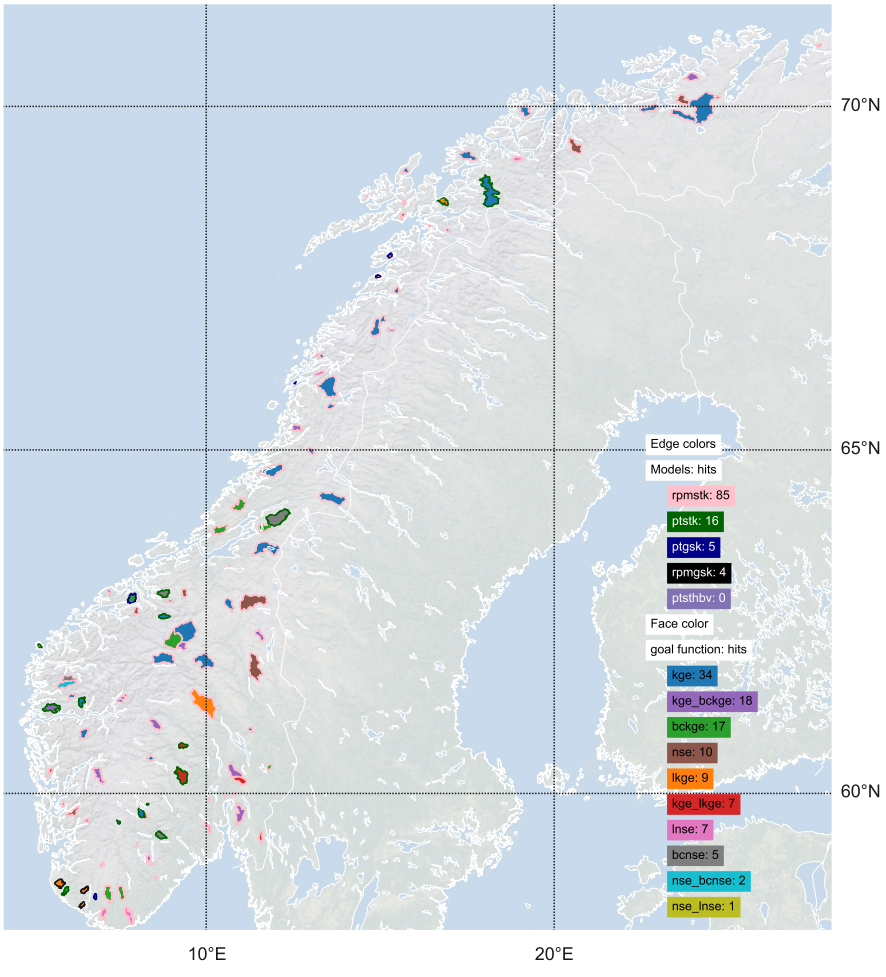

Figure 10. Spatial distribution of the best KGE scores without precipitation correction. The edge color of each of the catchments corresponds to the best model, the face color corresponds to the best goal function.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

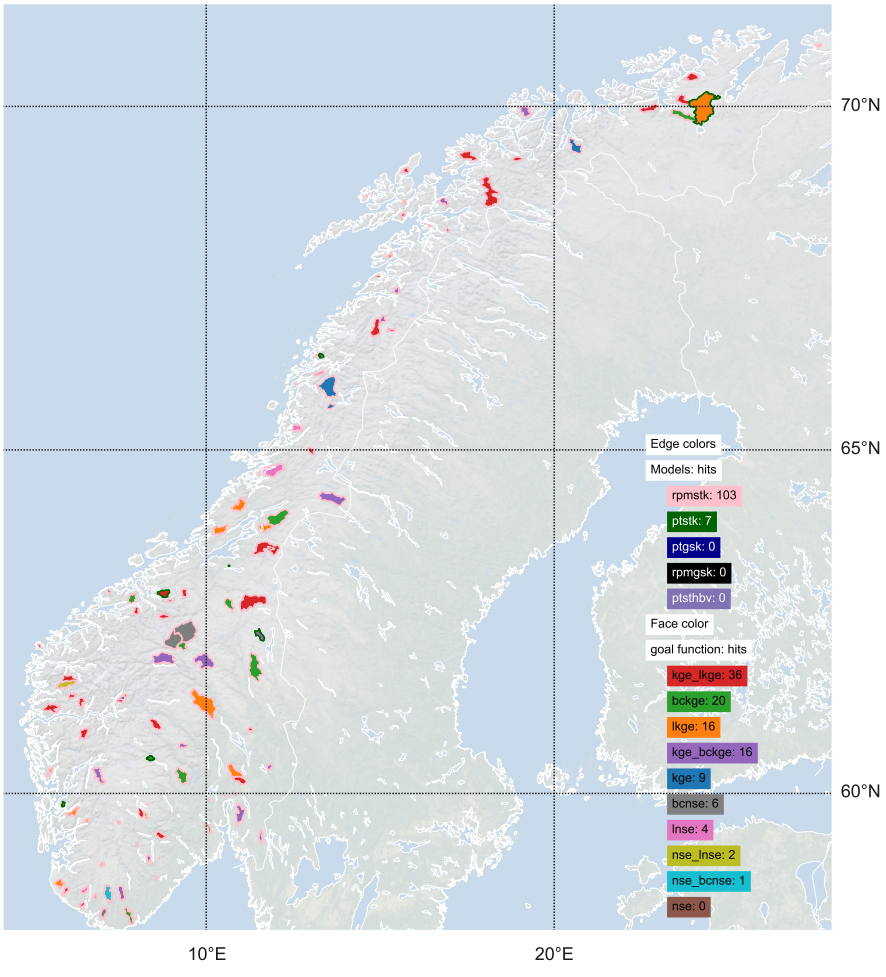

Figure 11. Spatial distribution of the best results with precipitation correction. The edge color of each of the catchments corresponds to the best model, the face color corresponds to the best goal function.

5 Discussion

5.1 Sensitivity to goal function selection

Santos et al. (2018) showed that KGE criteria calculated on log-transformed flows has pitfalls, such as oversensitivity near-zero flows, dependence on units and unintended excessive weight on the low flows compared to more balanced KGE. Thus,

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

for example, Thirel et al. (2023) excluded log transformed KGE from analysis. Our simulation results confirm finding of Santos et al. (2018). The LKGE performance has lowest median value for all stack configurations, followed by LNSE, if no precipitation correction is applied. If precipitaiton is corrected, the LNSE gets lowest med(KGE), followed by LKGE and bcNSE. The combination (KGE_bcKGE) is a transformation with minimal number of outliers in the catchment population and highest number of runs satisfying three criteria (NSE>0.0, KGE>-0.41 and |PBIAS|<15%) with precipitation correction. Without precipitation correction, this target also performs well, being only slightly behind KGE in three criteria metric. As Santos et al. (2018) mentioned in their discussion: *there is no ideal solution to avoid all problems*. The selection of the goal function in optimisation is very subjective, but different options maybe used to analyze parameter uncertainty in the models, (Cinkus et al., 2023; Onyutha, 2024). We showed that NSE-family of goal functions generate higher |PBIAS| and should be avoided in favor for KGE-based options. As expected, choice of KGE (and its variants except for LKGE discussed above) as the goal function produces higher KGE scores, whereas calibration against NSE (and its combinations) improves NSE scores. The high correlation between KGE and PBIAS is related to generally better PBIAS outputs of the models calibrated against KGE-based goal functions.

5.2 Sensitivity to model configurations and precipitation correction

Knoben et al. (2020) suggested that the seasonal benchmark might be a way to provide context for KGE scores interpretation better than -0.41 suggested before in Knoben et al. (2019b). In the study by Towler et al. (2023) focused on US catchments the mean flow performed better than median flow, with med(KGE) value of 0.08 and -0.1, respectively. Our results show that in Norwegian catchments med(Q) is better predictor than mean(Q), highlighting regional differences. However, med(KGE) for both mean and median flow benchmarks remain relatively low -0.01 and 0.06, respectively. These median values are slightly higher than -0.41 benchmark, but still give models a lot of room for beating the mark. As we have shown, the four out of five models used in this study outperform the seasonal benchmarks without precipitation correction, even though the processes representations inside each of them are different. One of the models (**PTSTHBV**) has a very narrow interquartile range of positive KGE scores, but was not able to beat the benchmark in some of the catchments. This model has a different routine for soil and runoff calculation. In addition, the **PTSTHBV** model, even though showed positive NSE values, revealed extremely low PBIAS values without precipitation correction. Interestingly, that **PTGSK** and **RPMGSK** demonstrated the best performance in terms of PBIAS metric without precipitaiton correction, but got poor NSE scores. These two models have highest spread of results between the models with some significant amount of poor performing catchments, but also containing well performing ones.

Even though, seNorge2018 is considered a high quality dataset (Erlandsen et al., 2021), the precipitation correction factor plays a significant role in improving overall models scores. This is partly due to runoff coefficient larger than 1 in 33% of catchments. Impact of precipitation correction is especially pronounced for the **PTSTHBV** model, see purple curve on the Fig. 5 compared to the Fig. 8. This model is hypersensitive to precipitation correction and has to be further scrutinized to find out reasons. Relatively low performing in terms of KGE and NSE scores **PTGSK** and **RPMGSK** models showed minimal responsiveness to precipitation correction. However, the two models have best performance in PBIAS metric if no precipitation

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

correction is applied, but this metric even deteriorates with precipitation correction. The **PTGSK** model was used in some of the previous studies, showing good NSE scores (Teweldebrhan et al., 2018; Matt et al., 2018; Bhattarai et al., 2020a, b). One can also take a look into zenode archive (Silantyeva and Huang, 2025) to find out that for some of the well performing catchments, the precipitation correction actually leads to improved scores, but the picture is somewhat mixed. Further research is needed to explore the underlying causes of the high variability and low median values of KGE and NSE scores for **PTGSK** and **RPMGSK** models.

The low performance of the **-GSK** stacks (**PTGSK** and **RPMGSK**) is especially pronounced for the Mountain and Inland hydrological regimes, see Fig. 9 and Fig. A2. These regimes are characterised by presence of winter or early spring low flows and at least one high flow period connected to snowmelt. The two models have gamma-snow semi-physics based snow model. The model is sensitive to the parameter, corresponding to the winter end day of the year, which was not calibrated in this study. Previous study by Newman et al. (2017) also showed that more simple conceptual model outperforms physically based counterpart. The limited sample size in the other catchment groups (Atlantic, Baltic, Transient) may have contributed to the observed consistency (narrower IQR) and better KGE and NSE performance for the two models. For the Atlantic regime the low flow happens during summer or early autumn and the high flow is due to the rainfall. The two models show good performance for this hydrological regime. This finding is inline with studies by Bhattarai et al. (2020a, b) for catchments in Nepal, where rainfall driven streamflow dominates. The PBIAS values and the impact of precipitation correction factor on well-performing **-STK** stacks (**PTSTK** and **RPMSTK**) suggest that for the Mountain and Inland hydrological regimes the quality of precipitation forcing remains an issue. Uncertainty from the precipitation forcing has to be further studied with the help of large-sample hydrology. Interestingly, for the Transient regime the best performing models (**-STK**) react minimally to precipitation correction, with even deteriorated PBIAS.

Huang et al. (2019) showed improved simulations with HBV model, where Penman-Monteith equation was used for evapotranspiration. For shyft.hydrology the **RPMSTK** model, which has Penman-Monteith routine, also tends to slightly outperform **PTSTK** model, which uses Priestly-Taylor equation for evapotranspiration.

5.3 Shortcomings and future work

In this study we used high resolution TIN-mesh. This limited our selection to the catchments with area < 1000 km^2. Future work should consider mesh of different shape and resolution and include bigger catchments. Our previous studies demonstrated improvements (Bhattarai et al., 2020b) and deterioration (Silantyeva et al., 2023) of performance scores with TIN-based simulations.

The split-sample test used in our work is assuming that the relationship between precipitation (P) and discharge (Q) remain stationary, which limits generalization to other conditions. To better assess model's abilities and avoid the influence of the nonstationarity the calibration and validation period could have been split in a different way, for example, using differential split-sample test (Li et al., 2012), or calibration on the full dataset or using odd-even approach (Arsenault et al., 2018). This limitation is subject for future analysis.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

As shown by Bárdossy et al. (2022) precipitation uncertainty can solely be responsible for up to 50% of model error. Our results show that precipitation correction plays significant role in improving model simulations. This is in line with previous studies with HBV model in the region. For example, Erlandsen et al. (2021) suggests precipitation correction coefficient between 0.5 and 3.0 for the seNorge2018 dataset. Huang et al. (2019) calibrated distributed HBV model with precipitation correction for undercatch in the range [0.5, 1.5]. The future work should consider comparing different forcings in the region and evaluate uncertainties related to each precipitation product.

The selection of stacks in this work is based on previous research and expert knowledge of the model. However, exploring sampling uncertainty may further improve selection process and help identify useful models, (Knoben et al., 2025). In addition, distinguishing models with structure, which closely matches dominant processes in the catchment, may help to reduce data related errors (Montanari and Di Baldassarre, 2013; McMillan et al., 2018).

This study focuses purely on streamflow simulation. However, the timing and the magnitude of snowmelt is an important part of hydrology in the snowmelt-dominated catchments. The future studies should evaluate the abilities of the flexible model configurations framework in the simulation of the snow-water equivalent (SWE) and the snow cover area (SCA).

6 Conclusion

Benchmarking hydrological model, such as Shyft is an important step towards quality simulations at regional scale. Shyft is a vital tool for several hydropower companies in Norway, helping forecast short and long term reservoir inflow and plan future hydropower project. To our knowledge, this is the first time, the flexible model configuration of Shyft is evaluated with the help of large-sample hydrology, providing valuable insights in the accuracy and robustness of the models. 109 catchments in mainland Norway were used for the evaluation. The seasonal benchmarks for the KGE score for the selected catchments are shown for the first time. The findings suggest that med(Q) is better predictor than mean(Q) in the population. We also introduce a combined metric (NSE> 0, KGE> -0.41 and $|\text{PBIAS}| < 15\%$) in our evaluation. In conclusion the following findings respond to the objectives identified in the introduction Section:

- The 5 evaluated model stacks are able to beat seasonal benchmarks most of the times, but different models have different capabilities. Stacks with temperature-index snow model and Kirchner runoff (**-STK**) are the most accurate and robust between the configurations. Stack, where HBV-soil and tank is used instead of Kirchner (**-HBV**), is oversensitive to precipitation correction, whereas stacks with minimalistic energy-balanced snow and Kirchner runoff (**-GSK**) show minimal responsiveness to precipitation correction.

- Among 10 evaluated goal functions KGE-based group is preferred over NSE-based group, except for the LKGE, which has known pitfalls. The KGE_bcKGE as the goal function, showed very promising results as a target goal function for all the models.

- The precipitation correction improves scores for 3 out of 5 stacks and for all goal functions.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

– Mountain and Inland hydrological regimes are the most sensitive to precipitation correction. Transient regime is the only hydrological regime, where precipitation correction does not improve performance scores. Atlantic regime is the only regime, where **-GSK** stacks perform similar to **-STK** based stacks.

All results of the study are publicly available for further analysis.

Code and data availability. The current version of the Shyft model is available from the project website at https://gitlab.com/shyft-os/shyft under the GPLv.3 license. A Zenodo archive with the results from this study available at https://doi.org/10.5281/zenodo.15595323. Discharge data, TINs and some of the scripts for final data processing and figure preparation for the publication available at https://gitlab.com/osilan/shyft-hydro-benchmarking

Author contributions. OS developed the study design. SS contributed to discussions and region specific analysis. OS developed the models and the simulation infrastructure utilized by the study. OS, SS and CYX discussed the study and prepared the manuscript.

Competing interests. No competing interests are present

Acknowledgements. This work is supported by Norwegian Research Council (NFR 336621) and contributes to Land-Atmosphere Interactions in Cold Environments (Latice) initiative at University of Oslo.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Allen, R. G., Trezza, R., and Tasumi, M.: Analytical integrated functions for daily solar radiation on slopes, Agricultural and Forest Meteorology, 139, 55–73, 2006.

Althoff, D. and Rodriguesa, L. N.: Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment, Journal of Hydrology, 600, https://doi.org/10.1016/j.jhydrol.2021.126674, 2021.

Arsenault, R., Brissette, F., and Martel, J.-L.: The hazards of split-sample validation in hydrological model calibration, Journal of Hydrology, 566, 346–362, https://doi.org/https://doi.org/10.1016/j.jhydrol.2018.09.027, 2018.

Bakke, S. J., Ionita, M., and Tallaksen, L. M.: The 2018 northern European hydrological drought and its drivers in a historical perspective, Hydrology and Earth System Sciences, 24, 5621–5653, https://doi.org/10.5194/hess-24-5621-2020, 2020.

Bárdossy, A., Kilsby, C., Birkinshaw, S., Wang, N., and Anwar, F.: Is Precipitation Responsible for the Most Hydrological Model Uncertainty?, Frontiers in Water, 4, https://doi.org/10.3389/frwa.2022.836554, 2022.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., and Wood, E. F.: Data Descriptor: Present and future Köppen-Geiger climate classification maps at 1-km resolution, Scientific Data, https://doi.org/10.1038/sdata.2018.214, 2018.

Bergstrom, S.: Principles and confidence in hydrological modelling, Nordic Hydrology, 22, 1991.

Bergström, S. and Lindström, G.: Interpretation of runoff processes in hydrological modelling—experience from the HBV approach, Hydrological Processes, 29, 3535–3545, https://doi.org/10.1002/hyp.10510, 2015.

Beven, K.: Benchmarking hydrological models for an uncertain future, Hydrological Processes, 14882, https://doi.org/10.1002/hyp.14882, 2023.

Bhattarai, B. C., Burkhart, J. F., Tallaksen, L. M., Xu, C.-Y., and Matt, F. N.: Evaluation of global forcing datasets for hydropower inflow simulation in Nepal, Hydrology Research, 51, 202–225, https://doi.org/10.2166/nh.2020.079, 2020a.

Bhattarai, B. C., Silantyeva, O., Teweldebrhan, A. T., Helset, S., Skavhaug, O., and Burkhart, J. F.: Impact of Catchment Discretization and Imputed Radiation on Model Response: A Case Study from Central Himalayan Catchment, Water, 12, https://doi.org/10.3390/w12092339, 2020b.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., de Amorim, P. B., Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Y., Chen, Y., Chifflard, P., Claps, P., Clark, M. P., Collins, A. L., Croke, B., Dathe, A., David, P. C., de Barros, F. P. J., de Rooij, G., Baldassarre, G. D., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Bevacqua, A. G., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T. H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M. L. R., Lindquist, E., Link, T., Liu, J., Loucks, D. P., Luce, C., Mahé, G., Makarieva,

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B. D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V. O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M. J., Post, D., Sierra, C. P., Ramos, M.-H., Renner, M., Reynolds, J. E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D. E., Rosso, R., Roy, T., Sá, J. H., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R. C., Skaugen, T., Smith, H., Spiessl, S. M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R. J., van der Ploeg, M., Loon, A. F. V., van Meerveld, I., van Nooijen, R., van Oel, P. R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A. J., Ward, P., Westerberg, I. K., White, C., Wood, E. F., Woods, R., Xu, Z., Yilmaz, K. K., and Zhang, Y.: Twenty-three unsolved problems in hydrology (UPH) – a community perspective, Hydrological Sciences Journal, 64, 1141–1158, https://doi.org/10.1080/02626667.2019.1620507, 2019.

Burkhart, J. F., Matt, F. N., Helset, S., Abdella, Y. S., Skavhaug, O., and Silantyeva, O.: Shyft v4.8: a framework for uncertainty assessment and distributed hydrologic modeling for operational hydrology, Geoscientific Model Development, 14, 821–842, https://doi.org/10.5194/gmd-14-821-2021, 2021.

Cinkus, G., Mazzilli, N., Jourde, H., Wunsch, A., Liesch, T., Ravbar, N., Chen, Z., and Goldscheider, N.: When best is the enemy of good – critical evaluation of performance criteria in hydrological models, Hydrology and Earth System Sciences, 27, https://doi.org/10.5194/hess-27-2397-2023, 2023.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resources Research, 51, 2498–2514, https://doi.org/https://doi.org/10.1002/2015WR017198, 2015.

Dingman, S. L.: Physical Hydrology, Waveland Press, Inc, 2015.

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

Erlandsen, H. B., Tallaksen, L. M., and Kristiansen, J.: Merits of novel high-resolution estimates and existing long-term estimates of humidity and incident radiation in a complex domain, Earth System Science Data, 11, https://doi.org/10.5194/essd-11-797-2019, 2019.

Erlandsen, H. B., Beldring, S., Eisner, S., Hisdal, H., Huang, S., and Tallaksen, L. M.: Constraining the HBV model for robust water balance assessments in a cold climate, Hydrology Research, https://doi.org/10.2166/nh.2021.132, 2021.

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, 2009.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, https://doi.org/10.5194/hess-18-463-2014, 2014.

Hanssen-Bauer, I. Førland, E., Haddeland, I., Hisdal, H., Mayer, S., Nesje, A. Nilsen, J., Sandven, S., Sandø, A., Sorteberg, A., and Ådlandsvik, B.: Climate in Norway 2100 – a knowledge base for climate adaptation, Tech. Rep. 1, The Norwegian Centre for Climate Services (NCCS), https://www.miljodirektoratet.no/globalassets/publikasjoner/M741/M741.pdf, 2017.

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

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

Huang, S., Eisner, S., Jan Olof Magnusson and, C. L., Yang, X., and Beldring, S.: Improvements of the spatially distributed hydrological modelling using the HBV model at 1 km resolution for Norway, Journal of Hydrology, 577, https://doi.org/10.1016/j.jhydrol.2019.03.051, 2019.

Kirchner, J. W.: Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, Water Resources Research, https://doi.org/10.1029/2008WR006912, 2009.

Knoben, W. J. M. and Spieler, D.: Teaching hydrological modelling: illustrating model structure uncertainty with a ready-to-use computational exercise, Hydrology and Earth System Sciences, 26, https://doi.org/10.5194/hess-26-3299-2022, 2022.

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geoscientific Model Development, 12, 2463–2480, https://doi.org/10.5194/gmd-12-2463-2019, 2019a.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019b.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments, Water Resources Research, 56, https://doi.org/10.1029/2019WR025975, 2020.

Knoben, W. J. M., Raman, A., Gründemann, G. J., Kumar, M., Pietroniro, A., Shen, C., Song, Y., Thébault, C., van Werkhoven, K., Wood, A. W., and Clark, M. P.: Technical note: How many models do we need to simulate hydrologic processes across large geographical domains?, Hydrology and Earth System Sciences, 29, 2361–2375, https://doi.org/10.5194/hess-29-2361-2025, 2025.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - a global community dataset for large-sample hydrology, Scientific Data, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Lawrence, D., Haddeland, I., and Langsholt, E.: Calibration of HBV hydrological models using PEST parameter estimation, Tech. Rep. 1, Norwegian Water Resources and Energy Directorate, 2009.

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., and Yan, D. H.: The transferability of hydrological models under nonstationary climatic conditions, Hydrology and Earth System Sciences, 16, 1239–1254, https://doi.org/10.5194/hess-16-1239-2012, 2012.

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

Matt, F. N., Burkhart, J. F., and Pietikäinen, J.-P.: Modelling hydrologic impacts of light absorbing aerosol deposition on snow at the catchment scale, Hydrology and Earth System Science, 22, 179–201, https://doi.org/10.5194/hess-22-179-2018, 2018.

McMillan, H. K., Westerberg, I. K., and Krueger, T.: Hydrological data uncertainty and its implications, WIREs Water, 2, https://doi.org/10.1002/wat2.1319, 2018.

Moges, E., Demissie, Y., Larsen, L., and Yassin, F.: Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis, Water, 13, https://doi.org/10.3390/w13010028, 2021.

Montanari, A. and Di Baldassarre, G.: Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty, Advances in Water Resources, 51, 498–504, https://doi.org/https://doi.org/10.1016/j.advwatres.2012.09.007, 35th Year Anniversary Issue, 2013.

Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Transactions of the ASABE, Soil and Water division, 50, https://doi.org/10.13031/2013.23153, 2007.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, 1970.

Newman, A. J., Mizukami, N., Clark, M. P., and Wood, A. W.: Benchmarking of a Physically Based Hydrologic Model, Journal of Hydrometeorology, 16, https://doi.org/10.1175/JHM-D-16-0284.1, 2017.

Nilsen, I. B., Hanssen-Bauer, I., Dyrrdal, A. V., Hisdal, H., Lawrence, D., Haddeland, I., and Wong, W. K.: From Climate Model Output to Actionable Climate Information in Norway, Frontiers in Climate, https://doi.org/10.3389/fclim.2022.866563, 2022.

Onyutha, C.: Pros and cons of various efficiency criteria for hydrological model performance evaluation, Proceedings of IAHS, 385, https://doi.org/10.5194/piahs-385-181-2024, 2024.

Powell, M.: The BOBYQA algorithm for bound constrained optimization without derivatives, Technical report damtp 2009/na06, Department of Applied Mathematics and Theoretical Physics, Cambridge, England, https://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf, 2009.

Priestley, C. and Taylor, R.: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, Monthly Weather Reiew, 100, 81–92, https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2, 1972.

Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, Journal of Hydrology, 11, https://doi.org/10.1016/j.jhydrol.2011.11.055, 2012.

Reistad, M., Breivik, Ø., Haakenstad, H., Aarnes, O. J., Furevik, B. R., and Bidlot, J.: A high-resolution hindcast of wind and waves for the North Sea, the Norwegian Sea, and the Barents Sea, Journal of Geophysical Research, 116, https://doi.org/10.1029/2010JC006402, 2011.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrology and Earth System Sciences, 22, https://doi.org/10.5194/hess-22-4583-2018, 2018.

Silantyeva, O. and Huang, S.: Supplementary files for "Benchmarking Shyft hydrologic model performance of streamflow simulations in mainland Norway" study (1.0), Zenodo, https://doi.org/10.5281/zenodo.15595323, 2025.

Silantyeva, O., Skavhaug, O., Bhattarai, B. C., Helset, S., Tallaksen, L. M., Nordaas, M., and Burkhart, J. F.: Shyft and Rasputin: a toolbox for hydrologic simulations on triangular irregular networks, preprint, https://doi.org/10.31223/X5CS95, 2023.

Skavang, J. Q.: Assessing the Shyft Modelling Framework in Nepal: Impact of Snow Routines and Terrain Representation on Simulated Water Balance Components, Master's thesis, University of Oslo, https://www.duo.uio.no/handle/10852/103795, 2023.

Tang, G., Clark, M. P., Knoben, W. J. M., Liu, H., Gharari, S., Arnal, L., Wood, A. W., Newman, A. J., Freer, J., and Papalexiou, S. M.: Uncertainty Hotspots in Global Hydrologic Modeling: The Impact of Precipitation and Temperature Forcings, BAMS, 24, https://doi.org/10.1175/BAMS-D-24-0007.s1, 2025.

Teweldebrhan, A. T., Burkhart, J. F., and Schuler, T. V.: Parameter uncertainty analysis for an operational hydrological model using residual-based and limits of acceptability approaches, Hydrology and Earth System Science, 22, 5021–5039, https://doi.org/10.5194/hess-22-5021-2018, 2018.

Thirel, G., Santos, L., Delaigue, O., and Perrin, C.: On the use of streamflow transformations for hydrological model calibration, egusphere, preprint, https://doi.org/10.5194/egusphere-2023-775, 2023.

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., and Zhang, Y.: Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States, Hydrology and Earth System Sciences, 27, https://doi.org/10.5194/hess-27-1809-2023, 2023.

Westergren, M.: Performance evaluation of regional calibration methods for a distributed hydrologic modelling framework, Master's thesis, University of Oslo, https://www.duo.uio.no/handle/10852/53238, 2016.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

615 www.ssb.no: https://www.ssb.no/en/energi-og-industri/energi/statistikk/elektrisitet, 2025.

Yang, X. and Huang, S.: Attribution assessment of hydrological trends and extremes to climate change for Northern high latitude catchments in Norway, Climate Change, 176, https://doi.org/10.1007/s10584-023-03615-z, 2023.

Yang, X., Yu, C., Li, X., Luo, J., Xie, J., and Zhou, B.: Comparison of the Calibrated Objective Functions for Low Flow Simulation in a Semi-Arid Catchment, Water, 14, https://doi.org/10.3390/w14172591, 2022.

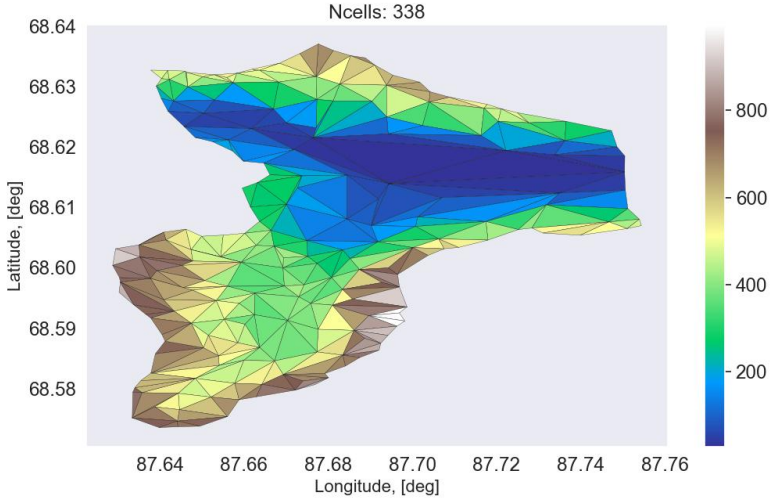

Figure A1. Example TIN-mesh for station 178.1.0 with app. area 36 km^2. the mesh contains 338 TIN-cells.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

Table A1: Model parameters and their min max values during calibration

Model					ptstk	rpmstk	ptgsk	rpmgsk	ptsthbv
module	parameter	value	min	max					
Radiation	albedo	0.25	-	-	o	x	o	x	o
	turbidity	1.0	-	-	o	x	o	x	o
Precipitation	pcorr	1.0*	0.4	2.0	x	x	x	x	x
Actual evaporation	ae_scale	1.0	-	-	x	x	x	x	x
Priestley-Taylor	albedo	0.2	-	-	x	o	x	o	x
	alpha	1.26	-	-	x	o	x	o	x
Penman-Monteith	height_veg	0.15	-	-	o	x	o	x	o
	rl	72	-	-	o	x	o	x	o
	full_model	0	-	-	o	x	o	x	o
Snow Tiles	tx	0.1	-8.0	0.0	x	x	o	o	x
	cx	0.33	-1.0	1.2	x	x	o	o	x
	ts	-0.12	-0.15	0.05	x	x	o	o	x
	lwmax	-0.12	-0.15	0.05	x	x	o	o	x
	cfr	-0.12	-0.15	0.05	x	x	o	o	x
Gamma Snow	tx	0.1	-2.0	2.0	o	o	x	x	o
	wind_scale	1.89	1.0	6.0	o	o	x	x	o
	FADR	6.1	5	18	o	o	x	x	o
	SADR	35.3	20	40	o	o	x	x	o
	max_water	0.1	-	-	o	o	x	x	o
	winter_EndDOY	100	-	-	o	o	x	x	o
	n_winter_days	221	-	-	o	o	x	x	o
	snowfall_reset_depth	5.12	-	-	o	o	x	x	o
	glacier_albedo	0.23	-	-	o	o	x	x	o
	wind_const	1.0	-	-	o	o	x	x	o
	surface_magnitude	30	-	-	o	o	x	x	o
	initial_bare_ground	0.04	-	-	o	o	x	x	o
	snow_cv_forest_factor	0.00	-	-	o	o	x	x	o
	snow_cv_altitude_factor	0.00	-	-	o	o	x	x	o
Kirchner	c1	-3.33	-8.0	0.0	x	x	x	x	o
	c2	0.33	-1.0	1.2	x	x	x	x	o

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

	c3	-0.12	-0.15	0.05	x	x	x	x	o
HBV-soil	soil.fc	250	50	500	o	o	o	o	x
	soil.lpdel	0.8485	-	-	o	o	o	o	x
	soil.beta	1.5	1	4	o	o	o	o	x
	soil.infmax	2	-	-	o	o	o	o	x
	tank.lpdel	0.8485	-	-	o	o	o	o	x
	tank.uz1	20.0	10	100	o	o	o	o	x
	tank.uz2	50.0	-	-	o	o	o	o	x
	tank.kuz0	0.05	-	-	o	o	o	o	x
	tank.kuz1	0.1	0.01	1.0	o	o	o	o	x
	tank.kuz2	0.5	0.1	1.0	o	o	o	o	x
	tank.perc	1.0	0.5	2.0	o	o	o	o	x
	tank.klz	0.05	0.01	0.1	o	o	o	o	x
	tank.ce	0.17 / 24.0	-	-	o	o	o	o	x
	tank.cevpl	1.1	-	-	o	o	o	o	x
Glacier Melt	dtf	1.0	-	-	x	x	x	x	x
	direct_response	0.0	-	-	x	x	x	x	x
Routing	velocity	0.0	-	-	x	x	x	x	x
	alpha	7.0	-	-	x	x	x	x	x
	beta	0.0	-	-	x	x	x	x	x

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

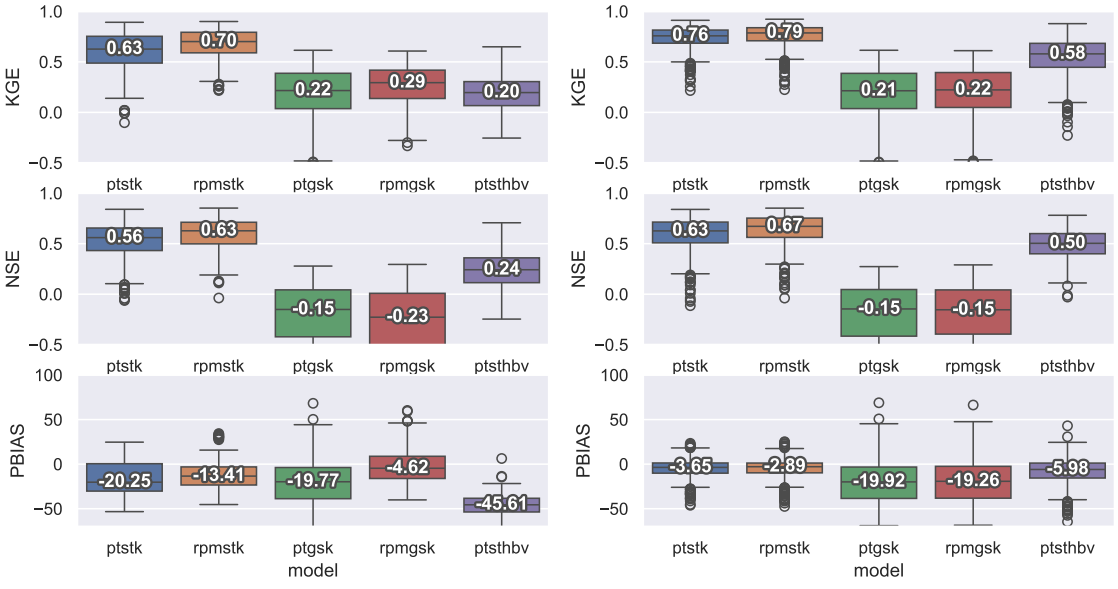

Figure A2. Kling-Gupta efficiency (KGE), Nash-Sutcliffe efficiency (NSE) and Percent Bias (PBIAS) scores based on daily streamflow for catchments classified as Inland. Performance scores are averaged for all catchments and 10 goal functions. Validation and calibration periods are considered separate runs. Left column: no precipitation correction; right column: with precipitation correction.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

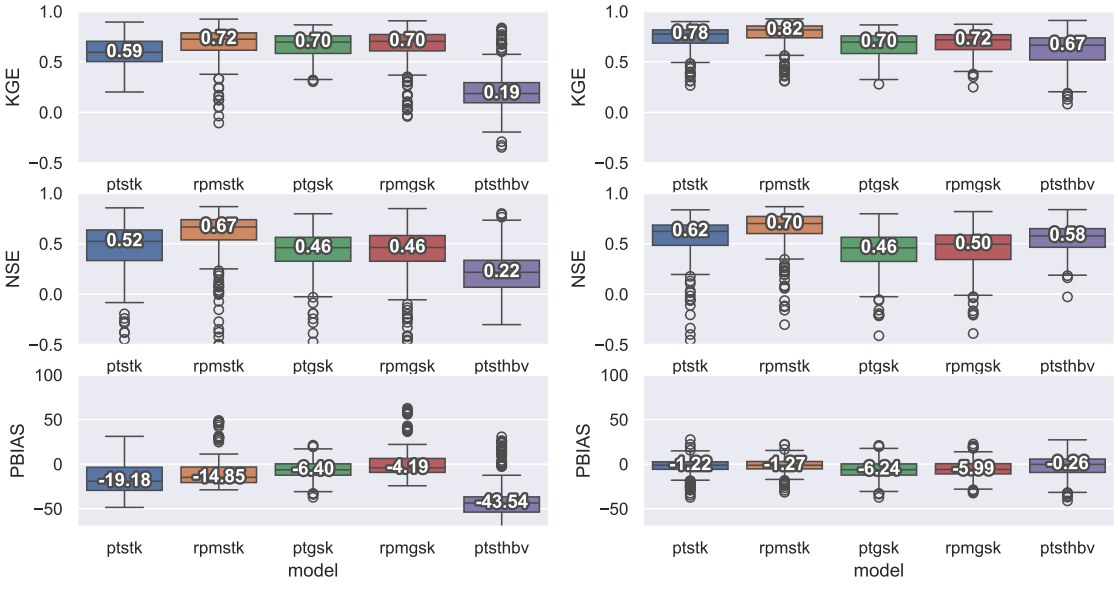

Figure A3. Kling-Gupta efficiency (KGE), Nash-Sutcliffe efficiency (NSE) and Percent Bias (PBIAS) scores based on daily streamflow for catchments classified as Atlantic. Performance scores are averaged for all catchments and 10 goal functions. Validation and calibration periods are considered separate runs. Left column: no precipitation correction; right column: with precipitation correction.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

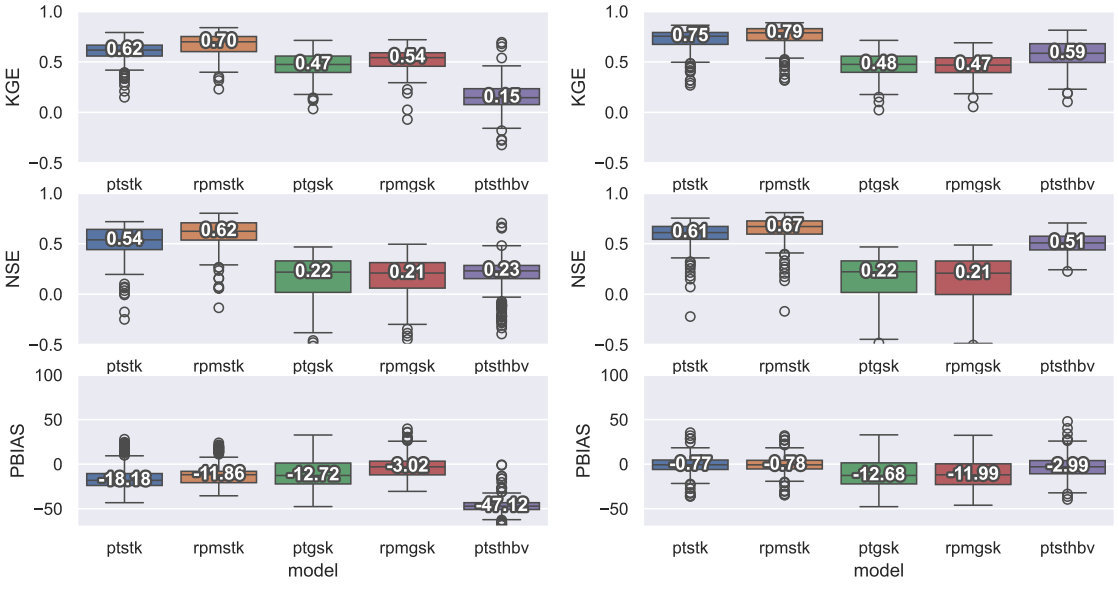

Figure A4. Kling-Gupta efficiency (KGE), Nash-Sutcliffe efficiency (NSE) and Percent Bias (PBIAS) scores based on daily streamflow for catchments classified as Baltic. Performance scores are averaged for all catchments and 10 goal functions. Validation and calibration periods are considered separate runs. Left column: no precipitation correction; right column: with precipitation correction.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

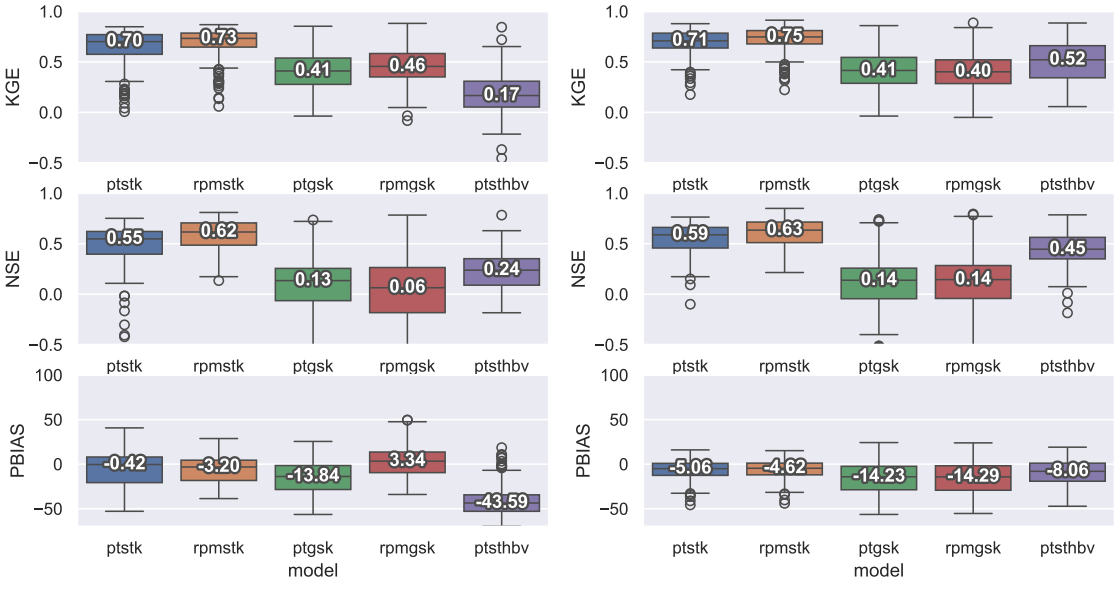

Figure A5. Kling-Gupta efficiency (KGE), Nash-Sutcliffe efficiency (NSE) and Percent Bias (PBIAS) scores based on daily streamflow for catchments classified as Transient. Performance scores are averaged for all catchments and 10 goal functions. Validation and calibration periods are considered separate runs. Left column: no precipitation correction; right column: with precipitation correction.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

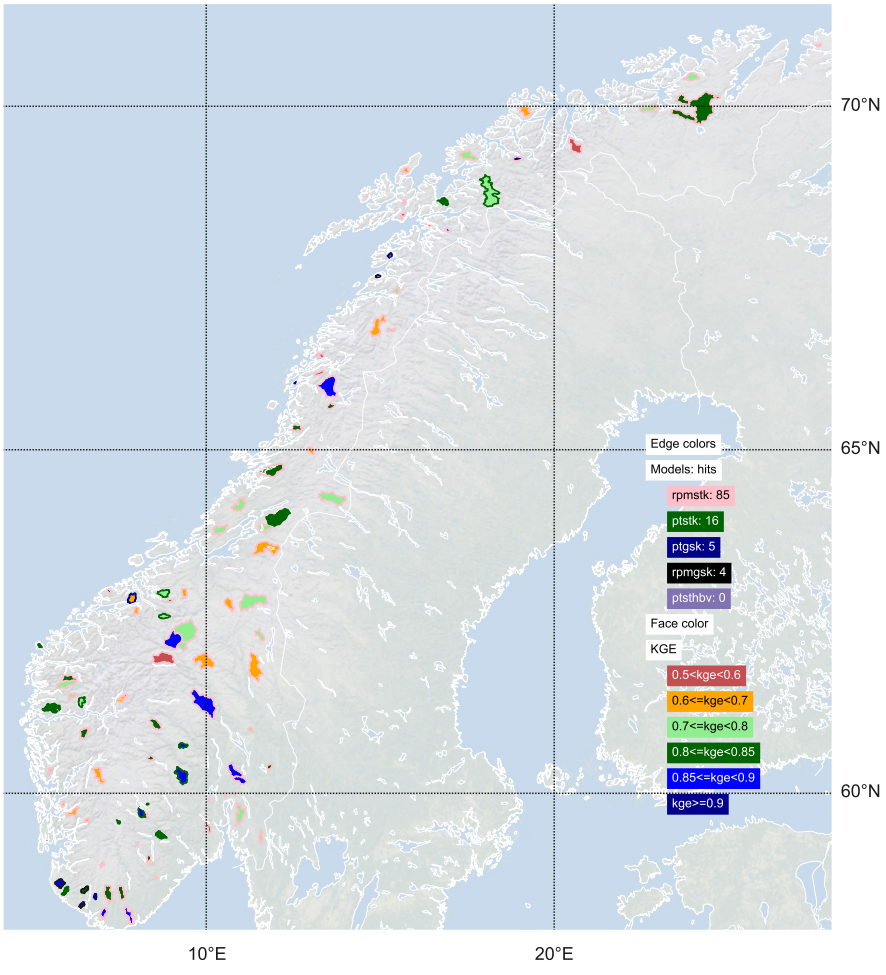

Figure A6. Spatial distribution of the KGE scores without precipitation correction.

https://doi.org/10.5194/egusphere-2025-4071
Preprint. Discussion started: 30 October 2025

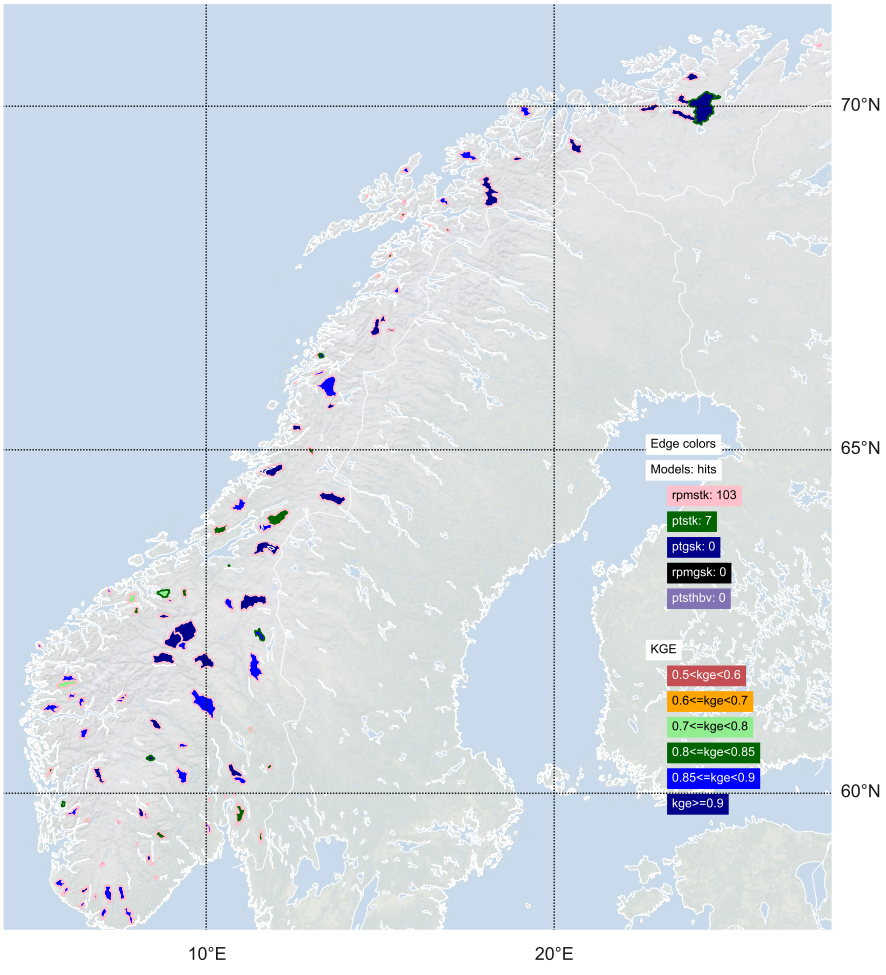

Figure A7. Spatial distribution of the KGE scores with precipitation correction.