

## Discussion

### 5.1 Main Added Value and Comparative Performance

This study provides the first public benchmark for the Shyft framework across 109 catchments in mainland Norway, directly addressing the significant underrepresentation of polar and alpine catchments in global hydrological research. Ruzzante et al. (2025) recently highlighted that while cold climates (including polar tundra) occupy a substantial portion of the earth's land mass, they account for a disproportionately small fraction of catchments in common large-sample dataset. By systematically exploring structure, calibration-objective and forcing choices we quantify how these three axes jointly determine simulation quality and operational suitability in seasonally snow-driven catchments.

By evaluating flexible "stacks" against seasonal climatological baselines, an ensemble of simple benchmarks (Knoben et al., 2025), we provide an objective context that moves beyond traditional, crude mean-flow benchmarks ( $NSE = 0$ ,  $KGE = -0.41$ ), which often do not impose sufficient constraints on models. While traditional guidelines (e.g., Moriasi et al., 2007) suggest static thresholds like  $NSE > 0.50$  for "satisfactory" performance, our results support the shift towards seasonal climatological benchmarks to provide a more demanding and objective context for model skill.

In the study by Towler et al. (2023) focused on US catchments, the annual mean flow ( $mean(Q)$ ) performed better than annual median flow ( $med(Q)$ ). In the Norwegian catchments  $mean(Q)$  is also better predictor than  $med(Q)$ . Among the simple benchmarks, daily mean flow is the best. As we have shown, all models used in this study outperform the annual benchmarks ( $mean(Q)$  and  $med(Q)$ ), even though the process representations inside each of them are different. However, only two (three) models outperform the daily mean flow benchmark without (with) precipitation correction, suggesting the high seasonality of the population (Knoben et al, 2025). The two models that are well performing against all simple benchmarks are also performing reasonably well when evaluated on the ( $KGE(1/Q)$  and interannual NSE. This suggests that – STK based models, besides being relatively simple in the structure, are robust for the variable hydroclimatic conditions evaluated in the study.

The selection of the goal function in optimisation is subjective, but different options may be used to analyze parameter uncertainty in the models (Cinkus et al., 2023; Onyutha, 2024) or focus on specific parts of the hydrograph (Thirel et al, 2023). Our findings are consistent with existing literature regarding the superiority of KGE-based objective functions for controlling water balance bias compared to NSE-based variants (Althoff and Rodrigues, 2021).

Furthermore, we confirm the documented numerical pitfalls of KGE, calculated on log-transformed flow (LKGE) (Santos et al., 2018), which generated a significant number of outliers

in our Norwegian catchments set, due to its oversensitivity to near-zero flow values. We also show that LNSE as a goal function variant, has similar optimisation issues as LKGE. Our results indicate that KGE calculated using Box–Cox transformation on the flow (bcKGE) and a combined metric KGE\_bcKGE represent promising, robust targets for high-latitude simulations, the latter is one of the best in satisfying three criteria: ( $NSE > 0.0$ ,  $KGE > -0.41$  and  $|PBIAS| < 15\%$ ). In addition, both bcKGE and KGE\_bcKGE are reasonably sensitive to low flows and perform average when evaluated on interannual NSE (Ruzzante et al. 2025).

While increasing model agility—specifically by including more sophisticated physics or more calibrated parameters—are expected to improve historical performance, our findings echo Newman et al. (2017): higher agility through complex parameterizations can lead to decreased transferability (such as with GS-based stack) and increased sensitivity to input quality (such as the case of –HBV based model). Furthermore, our results align with Knoben et al. (2020) in finding no systematic evidence that model complexity (parameter count) guarantees superior accuracy. In addition, we argue that in our case, model complexity (–GSK models) led to significant equifinality, such that local optimisation algorithms could not find global minima during calibration for many cases. The (–STK) based models not only have simpler equations with less parameters but also implement additional smoothing algorithms (Kawetski and Kuczera, 2007), which increases probability of reaching global minima.

## 5.2 Methodological Contributions: Flexible Frameworks and Precipitation Forcings

Methodologically, this study validates the utility of flexible modelling frameworks (similar to SUMMA and MARRMoT) for systematically exploring the "model hypothesis space". By assembling varied "stacks," we demonstrate a procedure applicable to other heterogeneous northern environments where no single model structure is universally suitable. We evaluate several configurations, distinguished by evapotranspiration, snow representation, and runoff response methods. We demonstrate that before addressing structural deficiencies of the models, the precipitation forcing should be inspected and a catchment or region-specific correction applied. By evaluating models on combinations of criteria, we reduce the risk of selecting models that perform well on one statistic and fail on the other.

As shown by Bárdossy et al. (2022), precipitation uncertainty can solely be responsible for up to 50% of model error. Even though, the seNorge2018 dataset utilized in this study is considered a high-quality dataset (Erlandsen et al., 2021), our results show that precipitation correction plays significant role in improving model simulations, enhancing performance for 89% of catchments. This is in line with previous studies with the HBV model in the region. For example, Erlandsen et al. (2021) suggest 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 under catch in the range [0.5, 1.5].

We demonstrated that precipitation correction has observable regional differences: models improve the most, when precipitation is corrected in the Mountain and Inland catchments, characterized by significant snowmelt related flow peak; simulations on the catchments in the low laying Transient and rain-dominated Atlantic regimes are less sensitive to precipitation correction. This supports the hypothesis that snow under catch is the main source of precipitation uncertainty.

Different model structures reveal various sensitivity to precipitation correction. The impact of precipitation correction is especially pronounced for the PTSTHBV model. This model has an HBV-like tank and response models, with more calibration parameters than Kirchner runoff of – STK stacks. The model is hypersensitive to precipitation correction and must be further scrutinized to find out reasons. Relatively low performing in terms of KGE and NSE scores PTGSK and RPMGSK models, containing semi-physical snow model, showed minimal responsiveness to precipitation correction. However, the two models have the best performance in PBIAS metric if no precipitation 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). Our results suggest that multi-criteria model evaluation can suggest more robust model structures. In our study, more complex structures (PTSTHBV, PTGSK, and RPMGSK) are underperforming, in all metrics, including KGE(1/Q) and interannual NSE metric. Thus, the three models should not be used in similar catchment populations. Thereas, PTSTK and RPMSTK, which represent a balance between simplicity of process representation and hassle-free calibration, are showing acceptable performance on all the metrics used in the study.

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

### 5.3 Practical Implications: Hydropower and Water Management

Practically, these results are essential for the efficiency of the Norwegian hydropower sector, where accurate inflow forecasts are vital for managing 90% of the country's electricity production. We identified that -STK based configurations (Snow Tiles and Kirchner runoff) are the most robust for this domain, if precipitation correction is applied to close the water balance.

The hypersensitivity of the PTSTHBV model to input quality indicates that hydropower operators using this structure must prioritize high-quality forcing data or risk significant volumetric biases. Conversely, the minimal responsiveness of -GSK models suggests their potential structural deficiency and increased model equifinality, that hinders applicability,

especially, with local optimisation algorithms. Ultimately, this benchmark enables decision-makers to select fit-for-purpose models.

## 5.5 Limitations

We acknowledge that the breadth of this study imposes certain limitations. Attempting to address model structure, goal function sensitivity, and input uncertainty simultaneously makes the objectives arguably too wide, potentially diluting the focus on specific process-based diagnostics.

In this study, we used high resolution TIN-mesh. This limited our selection to the catchments with area <1000 km<sup>2</sup>. Future work should consider meshing different shapes 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 nonstationary 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 an odd-even approach (Arsenault et al., 2018). This limitation is subject to future analysis.

The future work could 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 the 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 are 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).