

Conclusion

Benchmarking hydrological models such as Shyft is a step toward improving regional-scale streamflow simulations and establishing performance expectations for both operational and research applications. Shyft is widely used by hydrologists in Norway for short- and long-term reservoir inflow forecasting, yet its flexible model configurations have not previously been evaluated using a large-sample hydrology approach. This study provides the first such assessment across 109 catchments in mainland Norway, offering new insights into model accuracy, robustness, and sensitivity to calibration choices and precipitation uncertainty. We also present, for the first time, seasonal flow benchmarks for Norwegian catchments and introduce a combined metric to support transparent, multi-criteria evaluation.

Findings addressing RQ1: Across the five evaluated model stacks, most configurations outperform the seasonal benchmarks for the majority of catchments, but their capabilities differ significantly. Stacks combining a temperature-index snow model with a Kirchner runoff model (-STK) demonstrate the strongest and most robust performance, indicating good generality across regimes. Configurations using HBV soil and tank components (-HBV) show high sensitivity to precipitation correction, while stacks with a minimal energy-balance snow model and Kirchner runoff (-GSK) show limited responsiveness to precipitation correction, suggesting structural constraints or a high degree of equifinality that reduces the optimiser's ability to converge on an optimal solution. Differences in the evapotranspiration routine (PT vs RPM) have minimal impact on performance overall; however, within the -STK models, the RPM stack tends to outperform the PT stack.

Findings addressing RQ2: KGE-based objective functions outperform NSE-based functions across structures, suggesting parameter sets that generalise better across seasons and regimes. The KGE_bcKGE objective emerges as a promising compromise between sensitivity to flow dynamics and control of equifinality. LKGE confirms known issues and should be used with caution, but combined metrics such as KGE_bcKGE and KGE_LKGE offer a balance between overall fit and a low-flow-centric objective.

Findings addressing RQ3: Precipitation correction improves performance for three out of five structures across all calibration-objective setups. However, the magnitude of improvement varies strongly across hydrological regimes. Mountain and Inland catchments exhibit the highest sensitivity, indicating substantial snow undercatch. The Transient regime remains largely unaffected, suggesting higher-quality precipitation forcing. The similarity between -STK and -GSK model performance in the rain-dominated Atlantic regime highlights the quality of precipitation data there, reinforcing

that snow undercatch is the main source of precipitation uncertainty in Mountain and Inland regimes.

Findings addressing RQ4: Daily mean flow is the best predictor among the HydroBM flow benchmarks. These new benchmarks help distinguish between model structures and lay the foundation for future benchmarking studies in the region.

Overall, these findings underscore that regional-scale hydrological model performance emerges from a balance among structural adequacy, forcing uncertainty, and equifinality—a balance that flexible modelling frameworks must navigate explicitly. A large-sample hydrology perspective provides an effective way to disentangle these interactions and guide development toward more robust and transferable models.

All study results are publicly available for further analysis.