

# Introduction

One of the primary goals of hydrological modelling is to develop process understanding that supports the construction of models capable of physically realistic, accurate, and reliable simulations across diverse hydrological environments and climatic conditions (Gupta et al., 2014). Achieving this objective requires adopting a large-sample hydrology approach to enable robust assessments of model generality and reliability under changing conditions (Gupta et al., 2014). Benchmarking a hydrological model—an exercise that evaluates its applicability for various purposes—has become an emerging trend within the hydrological modelling community (Beven, 2023; Newman et al., 2017; Knoben et al., 2020; Towler et al., 2023). Progress in large-sample hydrology and model benchmarking has been further supported by the development of datasets such as CAMELS for the United States (Addor et al., 2017) and, more recently, the global Caravan dataset (Kratzert et al., 2023).

Hydrological modelling is subject to multiple sources of uncertainty related to input data, calibration, model structure (Moges et al., 2021), and sampling (Knoben et al., 2025). Indeed, quantifying uncertainty in hydrological modelling has been identified as one of the unsolved problems in hydrology (Blöschl et al., 2019). Acknowledging input-data uncertainty is an important component of hydrological analysis and water management (McMillan et al., 2018). Some studies focus on precipitation uncertainty (Bárdossy et al., 2022), while others also incorporate temperature uncertainty (Engeland et al., 2016; Tang et al., 2025). The choice of an objective function for calibration contributes to parameter uncertainty (Onyutha, 2024), and multi-objective analysis is one possible approach to address calibration uncertainty (Moges et al., 2021). Model-structure uncertainty has recently received more attention, as it remains underrepresented in the training of future hydrologists (Knoben and Spieler, 2022). Knoben et al. (2025) discuss performance equivalence among models and introduce the concept of sampling uncertainty, which quantifies uncertainty associated with the selection of the period used to compute performance metrics.

Acknowledging this complexity—further amplified by the spatial and temporal variability of climates and catchment characteristics—hydrological modelling faces a major bottleneck: no single model is universally applicable to all catchments. Flexible model configurations explicitly aim to overcome this limitation. Frameworks such as the Structure for Unifying Multiple Modeling Alternatives (SUMMA) (Clark et al., 2015), the Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) (Knoben et al., 2019a), and Shyft (Burkhart et al., 2020) provide a configurable environment in which models can be constructed based on an understanding of underlying processes at

varying levels of complexity, enabling comprehensive analysis of model-structure uncertainty.

Despite substantial progress in large-sample hydrological research and the development of multi-fidelity modelling frameworks that support different levels of model complexity, several important knowledge gaps remain in our understanding of how model structure, calibration strategy, and input-data uncertainty jointly influence model performance at regional scales: a. Flexible modelling frameworks often permit many structural alternatives, but it remains unclear which structures are most transferable and robust across contrasting hydroclimatic conditions (Knoben et al., 2025), and how structural choices interact with precipitation uncertainty. b. The selection of a calibration objective function is known to affect parameter estimates and predictive skill, yet systematic, large-sample evaluations of alternative single- and multi-objective functions remain limited in research. c. Precipitation remains a dominant source of uncertainty in mountainous environments. However, the extent to which different model structures and hydrological regimes can benefit from precipitation correction has not been quantified at scale. d. While global benchmarks exist, alpine and polar catchments are underrepresented in large datasets (Ruzzante et al., 2025), underscoring the need to establish a foundation for model evaluation and create interpretable benchmarks for these catchments.

Norway, situated in the northern high latitudes, is particularly vulnerable to climate change and has experienced some of the strongest warming globally since 1980 (Hanssen-Bauer et al., 2017; Nilsen et al., 2022; Yang and Huang, 2023). Accurate forecasting of streamflow, floods, and hydropower inflow is critically important in Norwegian catchments, given the country's reliance on hydropower and its growing exposure to climate risks (Hanssen-Bauer et al., 2017; Nilsen et al., 2022). Recent studies have 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), and exploring regional trends and extremes (Yang and Huang, 2023).

Using Shyft—an open-source, fully FAIR (findable, accessible, interoperable, and reusable) framework for uncertainty analysis and hydrological modelling—as an example of a flexible modelling environment, this study addresses the following research questions motivated by the identified knowledge gaps:

- RQ1. How do alternative model structures perform across a large, hydroclimatically diverse set of Norwegian catchments, and which configurations offer the greatest robustness and generality?

- RQ2. How does the choice of calibration objective function influence performance and robustness across models and regimes?
- RQ3. To what extent does precipitation correction improve model performance, and how does this sensitivity vary across hydrological regimes and model structures?
- RQ4. Can regional and seasonal benchmarks be defined to support future model evaluation and provide a transparent basis for model comparison in Norway?

The remainder of the paper is organized as follows. Section 2 describes the study area and forcing data. Section 3 presents the hydrological model, performance metrics, and experimental design. Section 4 reports the simulation results. Section 5 discusses the findings, limitations, and directions for future work. Section 6 provides the conclusions.