

Response to the Second Review of “Evaluating Flexible Configurations of the Shyft Hydrologic Model Framework Across Mainland Norway” manuscript.

Reviewer text:

The article compares five hydrological model structures on a large set of catchments in Norway. Models are tested with ten alternative calibration functions and evaluated with three efficiency criteria. Two versions of the models are evaluated, without and with precipitation correction factor.

Though large-sample hydrology studies often give insights on modelling performance, I found the originality of this study is unclear. The sensitivity of model results to calibration options has been already analysed in the literature, with recent studies cited by the authors. The individual evaluation of models is useful but it is difficult to get general conclusions from this work given some limitations in the testing framework (see detailed comments below).

Several choices lack justifications and results should be analysed in more detail to get actual insights from this modelling experiment. Research questions should also be more clearly stated. This would help the authors better explaining and demonstrating the novelty of their work.

Answer:

We have reformulated knowledge gaps, research questions, and objectives. The revised version of Introduction is provided as an attachment to the Answer.

Reviewer text:

I found this study has too many weaknesses and should undergo in-depth modifications including strengthening the introduction and discussion, designing a more convincing testing framework and providing a more detailed analysis of results. I think this would make the article very different. That is why I advise rejection.

Answer:

Thanks to reviewer's comment, following which we have rewritten the Introduction, Discussion, and Conclusions sections. In particular, we have clearly stated the knowledge gaps, research questions, and objectives. The revised Introduction, Discussion and Conclusion are attached.

Answer to the "Detailed comments"

Reviewer text:

1. The abstract could be much more concise and to the point to highlight the main points. Notations (families of KGE and NSE criteria) are unclear without further explanations.

Answer:

We revised abstract to "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. We benchmark Shyft—an open-source, cross-platform framework—across 109 catchments in mainland Norway using daily simulations from 1981–2020. Five model configurations span key structural choices: two evapotranspiration formulations (Priestley–Taylor; Penman–Monteith), two runoff methods (Kirchner; HBV), and two snow modules (temperature-index; semi-physical). Each configuration is calibrated under ten objective functions drawn from the Kling-Gupta Efficiency (KGE) and Nash-Sutcliffe Efficiency (NSE), and evaluated with and without a precipitation-correction factor. Performance is compared using standard metrics (KGE, NSE, percent bias (PBIAS)), and climatological benchmarks; KGE components, KGE calculated on reciprocal flow ($KGE(1/q)$) and interannual NSE are reported in the Supplement. Among simple benchmarks daily mean is better flow predictor than any other variant. Best performing configurations outperform benchmarks for the majority of catchments. Models combining a temperature-index snow module with Kirchner runoff (-STK) deliver the strongest and most robust performance across

regimes; precipitation correction further reduces bias. Configurations with the semi-physical snow module (-GSK) show lower KGE/NSE—particularly in Mountain and Inland regimes—and limited responsiveness to precipitation correction, yet yield the lowest |PBIAS| without correction. The temperature-index snow plus HBV runoff model (-STHBV) is most sensitive to precipitation correction: it has the worst bias without correction but moves into the top three across all metrics when correction is applied. Across structures, KGE-based calibration objectives reduce bias more effectively than NSE-based ones, whereas log-transformed objectives produce unstable parameter sets and poor KGE scores. The study reveals the box-cox KGE (bcKGE) and combined KGE+bcKGE goal functions are possible “generalist” objectives, performing well on all the evaluation metrics. This large-sample benchmark clarifies how structure, calibration objective, and precipitation uncertainty jointly shape performance, provides flow benchmarks references for Norway, and supplies publicly available results to support method development and operational uptake.”

Reviewer text:

2. The introduction should better highlight the research questions addressed by the authors and explain what is the novelty of their study compared to existing works.

Answer:

We revised introduction to highlight knowledge gaps and state research questions and objectives. The revised introduction is attached.

Reviewer text:

3. Data: Given the problems noticed on water balance on a large part of the catchments, this issue should be analysed in more detail to better identify the possible causes. We understand that snow undercatch may be one of them, but this is unlikely to be the only one. Some Budyko-type framework may be useful to analyse problems between catchments. This would help better understand the modelling results.

Answer:

As we demonstrated for the rain-dominated catchments, the precipitation correction has very marginal effect on 4 out of 5 model structures. This supports the hypothesis, that the

Mountain catchments, seasonal and snow dominated, are more prone to precipitation undercatch. In addition, these catchments typically correspond to the areas less represented in the observational networks, which further increases uncertainty of forcings. We cite some other studies with similar forcing in the region (Erlandsen et al., 2021, Huang et al., 2019), which also showed a need for precipitation correction in order for hydrological models to maintain water balance. Additional analysis might be beneficial, but is not feasible in the timeframe of the response. We will look into this in our future studies.

Reviewer text:

4. Data: It is unclear whether there were gaps in flow time series. Streamflow data quality is not commented.

Answer:

The streamflow data was selected based on the criteria specified in the text: less than 5% of missing data. The streamflow data was quality controlled and the best available at the study time. The data is provided by the Norwegian Water and Energy Directorate (nve.no).

Reviewer text:

5. Performance metrics: The selected performance metrics (NSE and KGE) are biased towards high flows, so it is not a strong surprise that objective functions that should be preferred should also emphasize high flows to some extent.

Answer:

Thanks. In the manuscript, we emphasized by our combined metric that metrics like KGE_bckGE give more balanced results than pure KGE or NSE (lower number of outliers). We have now added KGE calculated on inverted flow KGE(1/q) to the evaluation to specifically emphasize low flows in the Supplementary attached with the response. As can be seen, bckGE, or KGE_bckGE appear to be more “generalist” goal functions, demonstrating minimal number of non-converged runs, acceptable performance in both high and low-flow metrics and acceptable PBIAS. The increased spread and higher values in the |PBIAS| metric is the main issue with NSE-based goal functions, LKGE and its combination as goal functions, though showing higher performance on the KGE(1/q) metric, have known pitfalls (Santos et al., 2018).

Reviewer text:

6. Objective functions: The authors cite a recent study which extensively tested various objective functions, with a similar (but more rigorous) testing protocol.

What does the study proposed here bring new? Besides, I was surprised that the authors ignore a recommendation from a previous study the cite not to use the **KGE criterion calculated on log-transformed KGE...** to come to the same conclusion that this criterion should not be used.

Answer:

The use of additional objective function was not particularly difficult in our set up, but it was an intention to check how sensitive the local calibration algorithm to the pitfalls described by Santos et al. (2018 Technical note). There was explicitly stated that: *“... even if it did not occur in our experiment, the issue described in this technical note may lead to problems during the calibration process. Indeed, it can create a strongly negative zone in the objective function hyperspace, which may negatively impact the performance of local calibration algorithms.”* We tested the statement and confirmed that LKGE and LNSE should be used carefully. Santos et al. (2018) only suggested to avoid LKGE, but we noticed that LNSE is also difficult for the optimization algorithm. In addition, we relate our work to the Thirel et al. (2024) *“In addition, the fact that it is difficult to identify a transformation leading to the best simulations overall or simply sufficiently high performance for the whole range of streamflow indicates that developing generic models fitting all purposes is still a challenging task for modellers.”* We selected a combined score as a possible approach to identify goal function or model which is most appropriate for the majority of catchments represented in the study. Thirel et al, 2024 showed that transformations focused too much on the specific flow regions (low or high flow specific transformations) are less “general”, suggesting some “generalist” transformations, demonstrating best average performance. Even though, our “combined metric” is much simpler than the rankings provided by Thirel et al. (2024), we show that the combinations like KGE_bcKGE as goal functions are suitable “generalist” objective for our region, though putting less weight on the low flows compared to pure box-cox KGE or LNSE.

Reviewer text:

7. Testing protocol: The testing protocol should be improved. First a full split sample test should be done on the two subperiods: calibration on first and validation on second and vice versa. Thus model evaluation in calibration and validation would be available on all data, which would help evaluating model robustness (comparison of mean performance on validation and mean performance on calibration) more rigorously. **Second**, the distribution of results should not mix calibration and validation results. Indeed this may favour more complex models which have more degrees of freedom in calibration. **Third it is** unclear if the local calibration tool used is well adapted to the level of complexity of the tested models, which may have more than ten parameters, with possible secondary optima in the response surface where the algorithm may be trapped. Last, only one-year warm-up is used. If there are catchments with strong groundwater contribution (I do not know if it is the case), a single year may not be sufficient. This choice should be justified.

Answer:

The test on different calibration and validation periods was not feasible in the timeframe of the response to the review. By the time of writing the simulations are not finished. All models in the study have good temporal transferability. We provide separate calibration and validation results in the Supplementary. Before selecting the local calibration algorithm bobyqa we performed several tests with other options like SCEU-A, and found it to be the best in the tradeoff between computational cost and accuracy. However, we understand that the more complex GS-based models can indeed generate a more “bumpy” optimization space that might lead the algorithm to stuck in the local minima instead of the global one. The more simple ST-based model not only one having less parameters, but also the one, having additional smoothing algorithm by Kawetski and Kuczera, 2007, which can improve the likelihood of reaching a global minima.

There is no significant ground flow contribution in our catchments. The simulations are performed considering hydrological year, and the warm-up also performed on the hydrological year.

Reviewer text:

8. The **-0.41 threshold** on KGE should be explained at first appearance in the text.

Answer:

We revised the text and added explanation for the threshold at the first appearance.

Reviewer text:

9. There is no information on parameter distributions obtained by the various models. Are they realistic? What can be learnt from these distributions on the limits of the five model structures? What can be learnt from the values of the precipitation correction factor on the water balance issues the catchment may face?

Answer:

We thank the reviewer for this important comment. We fully agree that analyzing and discussing the calibrated parameter distributions is essential for assessing the physical realism of the models, understanding the structural limitations of the five model structures, and interpreting the role of the precipitation correction factor in relation to catchment water balance.

In the original submission, we focused primarily on model performance metrics, and the calibrated parameter sets were not retained in a form suitable for post-analysis. As a result, the parameter distributions could not be analyzed at that stage. Following the reviewer's suggestion, we will rerun the full calibration and validation experiments, explicitly storing the final calibrated parameters for all model structures and catchments.

In the revised manuscript, we will:

Present the distributions of key calibrated parameters for each model structure; assess their realism and consistency with hydrological expectations; discuss what these distributions reveal about the strengths and limitations of the different model structures; analyze the values of the precipitation correction factor and their implications for water balance biases and potential data or process deficiencies at the catchment scale.

These additions will be included in a new subsection in Section 4.6, along with corresponding figures and discussion.

Reviewer text:

10. The comparison between the two modelling options on precipitation correction is a **bit repetitive** and direct comparison between graphs is difficult.

Answer:

We provide side by side comparison in the Supplementary attached to the current response.

Reviewer text:

11. It is unclear whether the correction factor is applied on solid precipitation only (to compensate for snow undercatch) or on all precipitations? Would it make a difference?

Answer:

The precipitation correction is applied to the total precipitation. However, we showed that snow-dominated and rain-dominated catchments behave differently, supporting the hypothesis of significant snow-undercatch in the first place.

Reviewer text:

12. The authors argued that models are robust but no result is shown on that.

Answer:

We demonstrated that all models have strong temporal transferability. In addition, we demonstrated that there is a group of models –STK based which maintain high level of performance across different climatic conditions of the catchments population and with all the evaluation metrics. We argue that –STK based models used with multi-objective multi-criteria evaluations are robust.

Reviewer text:

13. The authors do not really comment the possible complementarity between model structures. For example, if the best performing structure was selected (in calibration) for each catchment, what would be the gain (in validation) over each structure applied on all catchments individually?

Answer:

We thank the reviewer for this insightful comment. We agree that the potential complementarity between model structures across catchments is an important aspect that deserves further discussion.

To address this point, we will perform an additional analysis in which, for each catchment, the model structure achieving the best performance during the calibration period is selected and subsequently evaluated during the validation period. The validation performance obtained with this “catchment-specific best structure” is then compared with the validation performance obtained when applying a single model structure uniformly across all catchments.

The result will be analyzed, a dedicated paragraph will be added in Section 4.7 discussing these findings and their implications for model selection strategies in multi-catchment studies. In particular, we will highlight that (anticipatedly), while no single structure is universally optimal, calibration-based structure selection can partially exploit inter-catchment heterogeneity without introducing overfitting.

Reviewer text:

14. Be consistent with criteria notations in figures and in the text (lower case / upper case)

Answer:

Thank you for your patience. We revised the manuscript to ensure consistency.