

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

Reviewer text:

Summary

This paper investigates the functioning of the Shyft modelling framework in 109 basins across Norway. Shyft allows the user to construct hydrologic models from components. In this work, five different configurations are tested. The 5 models are calibrated for each of the 109 basins with 10 different objective functions (called goal functions in this paper). Model performance is then assessed through (1) CDFs (in which models are compared to two seasonal cycle benchmarks), (2) box plots of aggregated performance on KGE, NSE and PBIAS scores, and (3) maps showing which combinations of model and goal function give the highest KGE score in each of the 109 basins. This analysis is done twice, once while including a precipitation correction factor and once without this factor. The paper concludes with recommendations about (1) the type of Shyft options that work well in these Norwegian basins, (2) the preferential use of KGE over NSE for less biased simulations, and (3) the use of a precipitation correction factor.

Comments

I have made some line-by-line comments in the PDF. Below are some further thoughts that I could not easily fit into a specific comment in the PDF.

[1] The idea of running mosaics of different models in space (i.e., finding an appropriate model for each individual basin, not a single model that works well enough for all of them) has been around for a while, but actual implementations that test the validity of this approach is more of a recent development. In that sense, this paper is a timely contribution. However, the introduction might be improved by being a bit more specific about the current knowledge gaps and state-of-the-art for these approaches. It currently focuses heavily on Shyft, but does not expand much on the more general multi-model approach that Shyft allows a user to set up.

Response:

We thank the reviewer for his/her positive evaluation of our work. We agree with the general comment, following which we have rewritten the Introduction section (see separate file) and will improve other sections to emphasise the general lessons that can be learned from this benchmarking exercise.

Reviewer text:

[2] Related, the stated goal of this work (l. 73-75) is “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”. This suggest a general approach applied to a specific case, but after reading the work (and particularly the conclusions) my main takeaway is that we now have some idea which of the five tested Shyft configurations work well in Norway. It’s less clear to me which general lessons can be learned from this work that are valuable to readers who do not actively use Shyft to model Norwegian basins. The benchmarking seems fairly minimal to me (just a comparison of CDFs), the conclusions about objective functions are possibly not that surprising (see [3c] below), the accuracy of precipitation context seems quite dependent on the model (some models provide better simulations with P corrections, others do not and sometimes get worse), and the flow regimes are specific to Norwegian basins. It would be good to emphasize the more general aspects of the work to justify publication in a broad international journal like HESS, compared to publication in a more regional or model-focused journal.

Response:

This study demonstrates the need to evaluate flexible model configurations fairly using a large sample of catchments. Although we use a regional set of catchments and regionally developed models, the results reveal model deficiencies and highlight the critical importance of data quality. Benchmarking analysis should evaluate different model configurations to justify model applicability and reliability. The evaluation procedure is transferable to other regions. Many flexible-configuration models have not undergone benchmarking, which can limit confidence in their applicability. We will revise the paper to make the generality of our approach and findings explicit

Following the reviewer’s advice, we have rewritten the Introduction and Conclusion (attached as separate files) and will improve other related parts to explicitly state the generality of our approach and findings.

Reviewer text:

[3] I believe part of this is that the justification of various choices can be improved. Doing

so might help highlight what broader lessons might be learned from this work. See below.

[3a] Based on Figure 2 it seems to me that Shyft supports more model configurations than just the five tested in this work. Some justification about why specifically these five are tested and not others would be welcome. This should, I think, extend to describing why certain parameters are chosen for calibration and others are not: lines 398-399 for example suggest that one sensitive parameter was not calibrated at all. Being clearer about the reasoning for these choices would be good.

Response:

We address these points in our line-by-line responses. In brief, the choice of the five configurations was subjective but grounded in prior studies using this framework. Parameter selection for calibration was likewise informed by earlier work, as cited in the manuscript and reiterated below. Running all available configurations was infeasible in the timeframe of the project. We will consider other available configurations in the future.

Reviewer text:

[3b] This would also be a good opportunity to outline what can be learned from pairwise comparison of each of the models. For example, if I understand Table A1 correctly, the single difference between RPMSTK and RPMGSK is the choice of snow module (ST vs GS). This would mean that differences between both models can be more clearly attribute to specific modelling decisions. This is the core idea behind multi-hypothesis modelling frameworks such as SUMMA (which Shyft is apparently inspired by), and leaning more into this part of the literature (it's currently not particularly prevalent in the framing of the paper) could help emphasize the more broadly applicable conclusions from this work.

Response:

We agree that pairwise comparisons can help attribute performance differences to specific modelling choices, including snow routines (temperature-index vs semi-physics-based). However, this is not the only explanation; the GS code may contain hidden assumptions or errors. Both models are also heavily calibrated. The GS-based model's complexity can reduce the optimiser's flexibility and create a more "bumpy" objective space, increasing the risk of local minima. By contrast, the ST-

based model relies on simpler equations and includes additional smoothing (Kawetski and Kuczera, 2007), which can improve the likelihood of reaching a global minimum. A potential mitigation for GS-based optimisation issues is to run multiple initialisations; we did not pursue this here to align with prior studies and to manage computational cost, but we plan to test this in future work. To ensure fairness, we harmonised settings across models and used the same optimisation algorithm. We update manuscripts discussion section to reflect this.

Reviewer text:

[3c] The analysis broadly consists of looking at the performance of the models on either KGE alone (CDFs, maps) or on a combination of KGE, NSE and PBIAS. I understand that there must be some sort of consistent metric(s) to evaluate the 10 different objective/goal functions on, but it's not particularly clear to me why specifically these three (or this one) was chosen. For example, it's widely accepted that hydrologic models tend to perform best for the metric they were calibrated on, so it's possibly not very surprising that any goal function that uses KGE shows better performance on KGE than any of the NSE-family of goal functions do. I think that explaining a bit more clearly why the chosen analysis approach is a helpful thing to do would be good (e.g. connecting the methodology to existing research gaps, operational practice or something else).

Response:

Operational practice commonly relies on NSE, and many recent large-sample studies centre conclusions on NSE. As we show, relying solely on NSE can mask substantial PBIAS; hence, we included KGE. In the Supplement, we also provide KGE components and RMSE. We acknowledge that single-value metrics are insufficient to fully assess model quality (Ruzzante et al., 2025). To add context, we include seasonal benchmarks, and the updated Supplement documents the flow benchmarks most relevant to our region.

Reviewer text:

[4] Readability and clarity can be improved in my opinion. Most of my line-by-line comments focus on this. See also the points below.

Response:

We provide detailed line-by-line revisions below.

Reviewer text:

[4a] More generally, I struggled quite a bit keeping track of the five model acronyms. The letters don't mean much to me because I don't have the familiarity with the Shyft software to map the acronyms onto specific models. The acronyms are also rather similar and this makes the text quite hard to follow. I expect that even just naming the models "model 1 - 5" (or "stack 1 - 5" to stick with the Shyft terminology) would be easier to follow for readers.

Response:

Agreed. We will adopt simpler naming (e.g., "stack 1–5") in the main text for readability and provide a clear mapping to Shyft acronyms in the Supplement.

Reviewer text:

[4b] It's also not always clear if results in figures refer to the calibration or evaluation period. This should be clarified in all cases.

Response:

We found good temporal transferability across models, and combined plots improve readability. Final figures in the original manuscript show combined periods unless stated otherwise. In the updated manuscript we clearly split the calibration and validation periods. The Supplement provided with the response illustrates the way we plan to update the figures. The Zenodo archive includes separate calibration and validation results. We will clarify period coverage consistently in captions.

Answer to line-by-line comments:

1. Line 83: Fixed
2. Line 109: Fixed. Added: “Adding more catchments to underrepresented regimes was not possible due to our data requirements.” Quality-controlled data in Norway are limited; the selected set is the best feasible. Consequently, we could not add more catchments to underrepresented regimes.
3. Line 116. Revised “Shyft uses high ...” to “In the current study, Shyft uses a high ...”. This also addresses line 129. While Shyft supports various mesh configurations, we use TINs here.
4. Line 124. Fixed.
5. Line 127. Fixed. Thank you for your patience; we revised the text accordingly.
6. Line 129. Fixed, see 3.
7. Line 131-132. Revised to: “We test five model stacks (see Table A1): PTSTK, used operationally and in Skavang (2023) for a Nepalese catchment (PT = Priestley–Taylor evapotranspiration; ST = snow-tiles temperature-index model; K = Kirchner runoff response); RPMSTK (new; R = radiation correction; PM = Penman–Monteith evapotranspiration; -STK as above); PTGSK, previously best-performing operationally and evaluated for parameter uncertainty in the Nea–Nidelva catchment (Teweldebhran et al., 2018; GS = simple energy-balance snow routine); RPMGSK, which showed superior performance in a Nepalese catchment (Bhattarai et al., 2020b); and PTSTHBV (new), which replaces Kirchner with HBV soil and tank models.”
8. Line 135. Radiation correction is not tied to any specific snow model. Instead, it is part of the forcing-data processing workflow and can be enabled or disabled as needed. The text has been revised to clarify this point.
9. Line 136. Fixed.
10. Line 137. Model-configuration choices were guided by operational experience and, where applicable, prior research findings.
11. Line 156-157. Corrected. The notation is changed to the one used in Gupta et al. 2009: $\beta = \frac{\mu_s}{\mu_o}$; $\alpha = \frac{\sigma_s}{\sigma_o}$.
12. Line 164. Revised to: “... unlike NSE, it assigns equal weight to correlation, variability, and bias.”
13. Line 181-182. We did not implement special treatments for numerical issues known for LKGE and LNSE, but we verified their existence. Prior to full benchmarking, we tested zero-flow definitions and adopted a simple censoring ($q > 0$) for calibration. These preliminary results are provided in the Supplement.
14. Line 189-190. The authors were not aware of the Commentary from W. J. Knoben at the outset of the study, therefore, we followed the approach of E. Towler 2023, which employs climatological benchmarks. We contacted E. Towler to confirm that our interpretation of the benchmark and its calculations was correct, given unexpectedly low benchmark values. We now provide simple benchmarks calculated using HydroBM package for our dataset in the Supplementary. The model evaluated as “best” indeed beats all the hydroBM flow benchmarks. Our original discrepancy stems from using hydroGOF’s KGE (Kling et al., 2012), whereas HydroBM implements Gupta et al. (2009). The remaining KGE scores were calculated using the *hydroeval* python

package (KGE), which follows the formulation of Gupta et al. 2009. (Hallouin, T. (2021). *HydroEval: Streamflow Simulations Evaluator (Version 0.1.0)*. Zenodo. <https://doi.org/10.5281/zenodo.2591217>). We corrected our calculations and ensured consistency in the evaluation metric throughout the analysis.

15. Line 199-200. Non-converged runs were limited to a few catchments that are inherently challenging, as also identified in our model intercomparison project (submitted). Using multiple goal functions reduced non-convergence risk in this large-sample setting. We reported the “number of runs satisfying three criteria” to identify model/goal-function combinations with minimal non-convergence. In practice, regional experts would fine-tune models in specific catchments to achieve convergence.
16. Line 219-220. We introduced the three-criteria approach to identify model/goal-function combinations that perform satisfactorily in most cases. Sole reliance on NSE can yield larger biases and, in some catchments, performance worse than $NSE = 0$. Using KGE alone is reasonable (and aligns with the combined metric) but diverges from operational reliance on NSE and reduces comparability. Our combined approach balances these aims. As W.J. Knoben notes, one-value scores do not guarantee realism. To aid interpretation, we provide HydroBM metrics in the Supplement (see also comment 14). In addition, we provide CDF plots for the component of KGE score, RMSE and LKGE scores.
17. Line 224 Fixed.
18. Line 230. Revised to: “In most cases, streamflow is underestimated; in some cases, it is overestimated.”
19. Figure 5. Caption. We added “the validation period has darker colour” to complement “the calibration period has lighter colour.” We also verified our seasonal-benchmarks using HydroBM package in the Supplement.
20. Line 245. Revised to: “The PTSTHBV model is the only one with scores intersecting both med(Q) and mean(Q) benchmarks.”
21. Line 245-246. Added caption text for clarity.
22. Line 257. Agreed. We will present figures side by side in the revision.
23. Line 277. We will place precipitation-corrected and uncorrected results side by side. Initially, we separated them to preserve readability, as side-by-side plots reduce figure size.
24. Line 287. We could not quickly identify the issue with the PTSTHBV stack; this will be investigated further. As a recent development, the model may contain implementation deficiencies.
25. Line 303-304. PTGSK and RPMGSK are largely insensitive to precipitation correction, and performance can degrade with correction. PTGSK parameter choices followed Teweldebrhan et al. (2018). After running the simulations, we found that one sensitive parameter had been fixed. While not influential in the single-catchment study, it affects regional differences, so the model may not operate in an optimal range in many catchments. The semi-physical snow component can also induce non-linear responses (e.g., additional precipitation routed as liquid flow instead of snow, or enhanced evaporation). As noted earlier, the added complexity may challenge the optimiser, increasing susceptibility to local minima.
26. Figure 9. Because CDFs showed good temporal transferability, we aggregated results thereafter. The accompanying dataset provides calibration

and validation results separately. If preferred, we can show validation-only figures. We also add the number of catchments to regime captions. The added Supplement shows validation period only.

27. Line 307. We agree that differences in the Inland regime are less pronounced. The Mountain regime shows strong sensitivity to precipitation correction, consistent with precipitation-undercatch patterns. Note also that Mountain includes 43 catchments versus 27 in Inland; the larger sample may accentuate differences.
28. Line 334. Yes, indeed, the calibration algorithm does not navigate the search space efficiently when an additional calibration parameter (precipitation correction) is introduced. The starting point changes (e.g., an initial precipitation correction of 0.8 if the range is [0.4, 1.2]), which can immediately steer the search in the wrong direction. We believe that the added complexity of the model's snow component creates more "bumpy" search space, where calibration algorithm is prone to becoming trapped in local minima. This illustrates the case, in which increased model complexity does not lead to any improvements for the intended application - streamflow simulation.
29. Line 336. We appreciate the referee for the patience in interpreting model results. We will improve the figure based on the suggestion.
30. Line 339-341. We are revisiting the combined metrics, now including the validation period. Therefore, we do not anticipate any issues with the optimization algorithm (except for the -GSK based models, as discussed above). We will show validation-only figures.
31. Line 347-348. Revised to: "If evaluation is based on KGE alone, there is always a model-goal-function combination that improves scores relative to seasonal benchmarks. However, as shown by Ruzzante et al. (2025), one-value metrics such as NSE and KGE are insufficient to characterise performance, especially in highly seasonal catchments where seasonal benchmarks can be relatively high. Still, the best-selected model outperforms all HydroBM benchmarks. We will assess interannual variability in our future work."
32. Line 374-376. In the Supplementary, we provide heatmaps in addition to CDF plots to support the statement.
33. Line 384-385. The runoff coefficient exceeds 1 in 33% of catchments due to known precipitation-undercatch issues, especially for snow in the mountainous catchments, see, for example (e.g., Lussana, 2018). As shown here, Atlantic (rain-dominated) catchments require less correction than Mountainous catchments, supporting snow undercatch as the primary driver.
34. Line 398-399. As noted in the answer 25, the GS-based setup followed Teweldebrhan et al. (2018), where the parameter in question was not considered sensitive and was fixed using best-practice knowledge. Internal discussions with Shyft developers indicate that this parameter should be calibrated in large-sample studies. It controls the "last day of accumulation," after which melt begins, and contributes to the model's "semi-physical" character.
35. Line 457. We state that the version of Shyft used in this study is 21. The repository <https://gitlab.com/osilan/shyft-hydro-benchmarking> is open-access.