

Supplementary Materials. Evaluating Flexible Configurations of the Shyft Hydrologic Model Framework Across Mainland Norway

Olga Silantjeva¹, Shaochun Huang², and Chong-Yu Hu¹

¹Department of Geosciences, University of Oslo, Sem Sælands vei 1, Blindern, 0371 Oslo, Norway

²Norwegian Water Resources and Energy Directorate (NVE), Middelthuns gate 29, 0368 Oslo, Norway

Correspondence: Olga Silantjeva (olga.silantjeva@geo.uio.no)

. TEXT

1 Supplementary materials

1.1 Setting context for benchmarking. KGE scores.

Figure 1 demonstrates CDF plots for flow benchmarks calculated using hydroBM packaged (Knoben, 2024). The daily mean flow and monthly mean flow benchmarks are to the right of all other flow benchmarks, indicating their better performance. Annual median flow is the worst performing benchmark for our catchments. We provide results for the validation and calibration periods for the models with and without precipitation correction. -STK based models outperform flow benchmarks in both precipitation cases. PTSTHBV model outperforms flow benchmarks only if precipitation is corrected. Models with minimalistic energy-balance snow (-GS-based) struggle to beat the monthly and daily mean and median flow benchmarks.

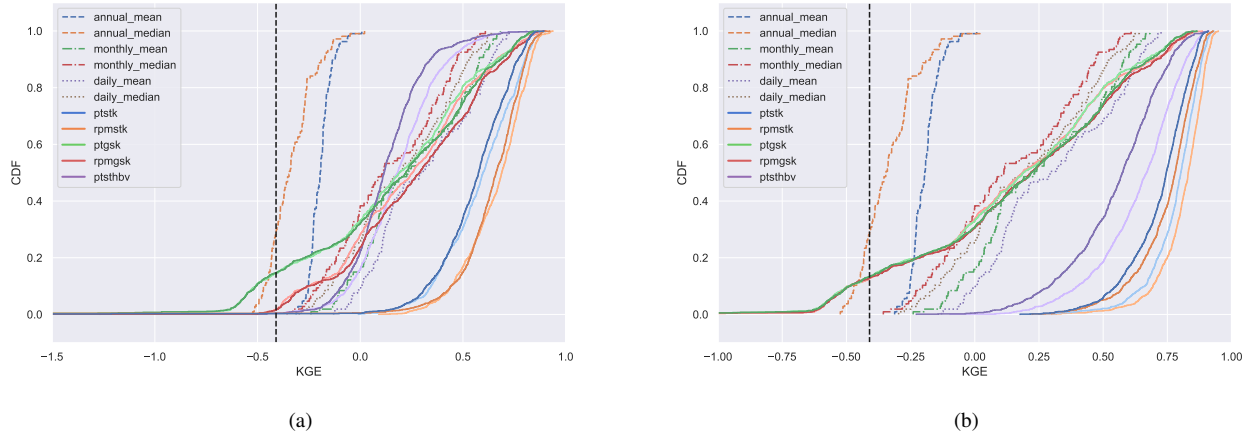


Figure 1. Cumulative density function plot of KGE scores for flow benchmarks calculated using hydroBM package (Knoben, 2024) and simulation results for models (a) without precipitation correction, (b) with precipitation correction. Validation period for the models is in the dark colors, calibration period – light colors

10 1.1.1 KGE components: α , β and r .

We present here components of KGE metric (Gupta et al., 2009). For all the plots in the subsection validation results are presented in darker colors than calibration. All models show strong temporal transferability.

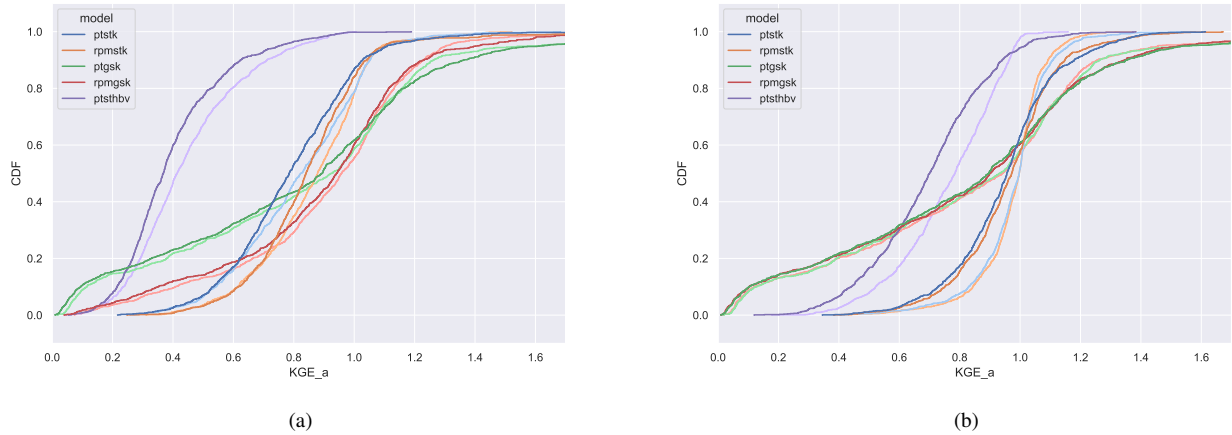


Figure 2. Cumulative density function plot of α component of KGE (Gupta et al. (2009)) simulation results for models (a) without precipitation correction, (b) with precipitation correction. Validation period for the models is in the dark colors, calibration period – light colors

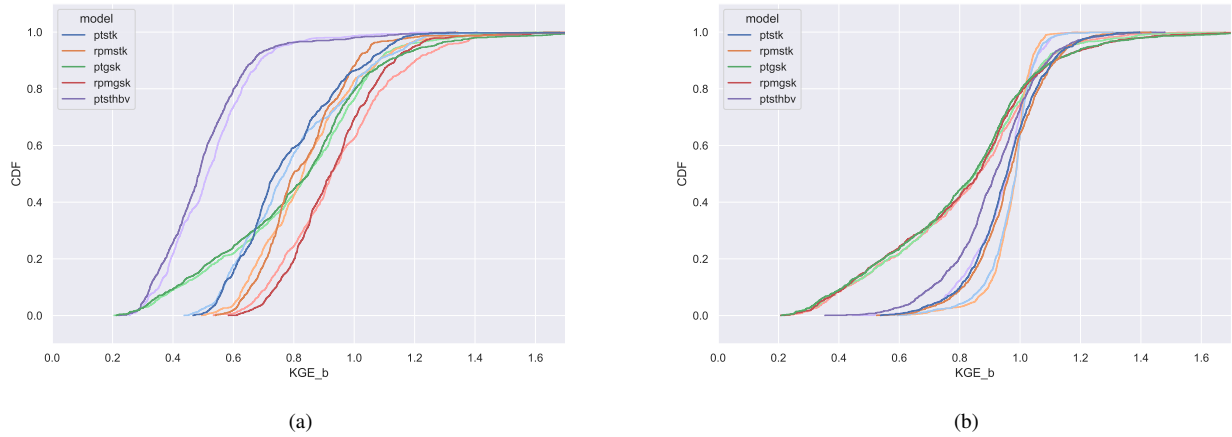


Figure 3. Cumulative density function plot of β component of KGE (Gupta et al. (2009)) simulation results for models (a) without precipitation correction, (b) with precipitation correction. Validation period for the models is in the dark colors, calibration period – light colors

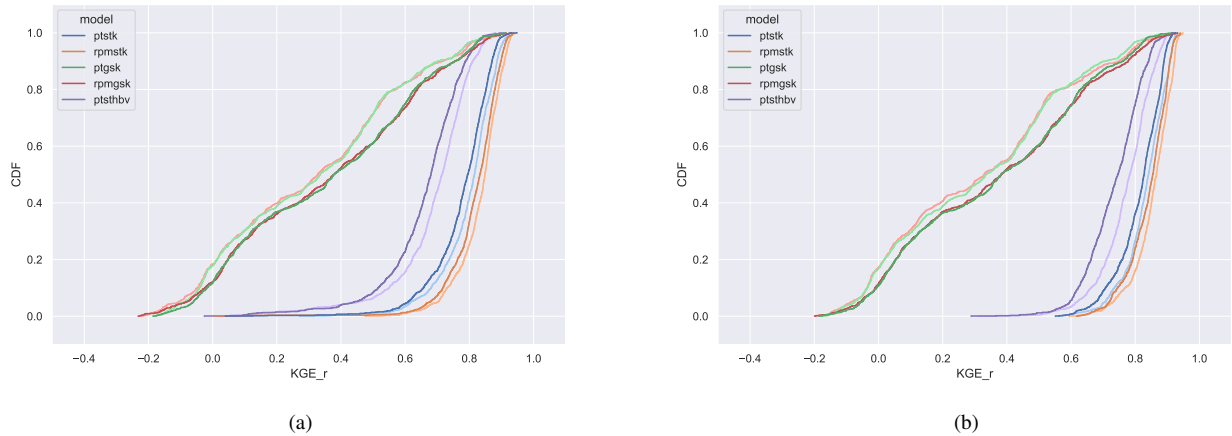


Figure 4. Cumulative density function plot of r component of KGE (Gupta et al. (2009)) simulation results for models (a) without precipitation correction, (b) with precipitation correction. Validation period for the models is in the dark colors, calibration period – light colors

1.2 Boxplots of performance metrics of the models. Validation period only.

The Fig. 5 demonstrate performance of the models on the KGE, NSE, PBIAS, KGE calculated on reciprocal flow ($KGE(1/q)$), which is suggested as a low-flow metric in Santos et al. (2018); Knoben et al. (2020); Thirel et al. (2023) and interannual NSE (Ruzzante et al. (2025)). The left panel correspond to the case without precipitation correction, the right – with precipitation

correction. All results are show for validation period only. KGE, NSE show strong performance of -STK based models, but precipitation has to be corrected to improve PBIAS. The three metrics are provided as complementary to the low-flow focused KGE(1/q) and interannual NSE.

20 As expected, precipitation correction has limited impact on the low flow or interannual variability metrics. The models which stand apart in both metrics are PTSTK and RPMSTK, which demonstrate relatively high values, thus suggesting their structures are adequate for all parts of hydrograph. The other structures struggle to reproduce low-flows well.

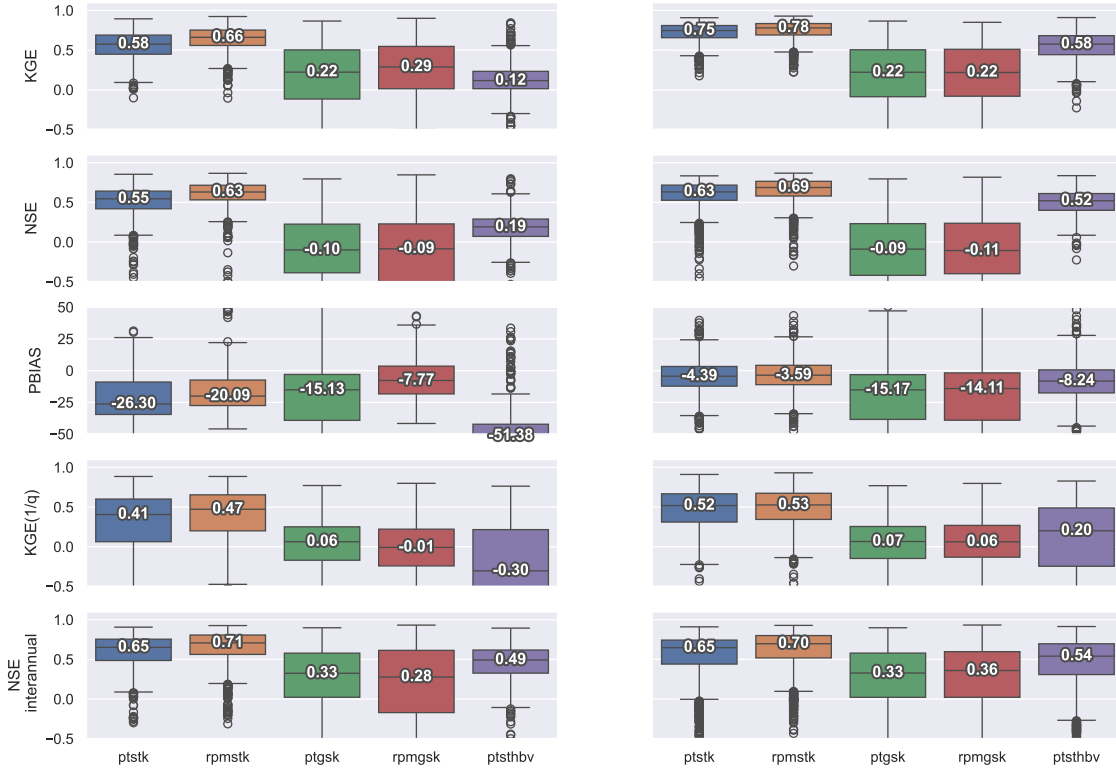


Figure 5. Boxplots for performance scores. Left panel: No precipitation correction, Right panel: with precipitation correction. From top to bottom: KGE, NSE, PBIAS, KGE(1/q), interannual NSE

1.3 Performance of Goal function. Validation period only.

The Fig. 6 demonstrates boxplots of scores for each goal function used in the study. On the left panel, we present results without precipitation correction; on the right panel: with precipitation correction. We show three metrics: KGE, NSE and PBIAS and

25

a combined metric: "Number of runs satisfying the three criteria". In addition, we present KGE calculated on the inverted flow ($KGE(1/q)$, kge_{inv} on the plots), which evaluates, how goal function choice affects low flow simulation. As can be seen, the log-transformed flows, as expected, perform better on this evaluation metric. But, the log-transformation on KGE has known pitfalls, see Santos et al. (2018); and we also found instability during calibration for log-transformed NSE. The interannual NSE (nseinter) Ruzzante et al. (2025) is shown at the bottom of the figure. All goal functions perform reasonably, when evaluated in this metric. If we identify a more "generalist" goal function, which takes into account not only low-flow or high-flow parts of the hydrograph, but overall water balance (PBIAS) and inter-annual variability, than the plot reveals bcKGE and KGE_bcKGE as a reasonable choice.

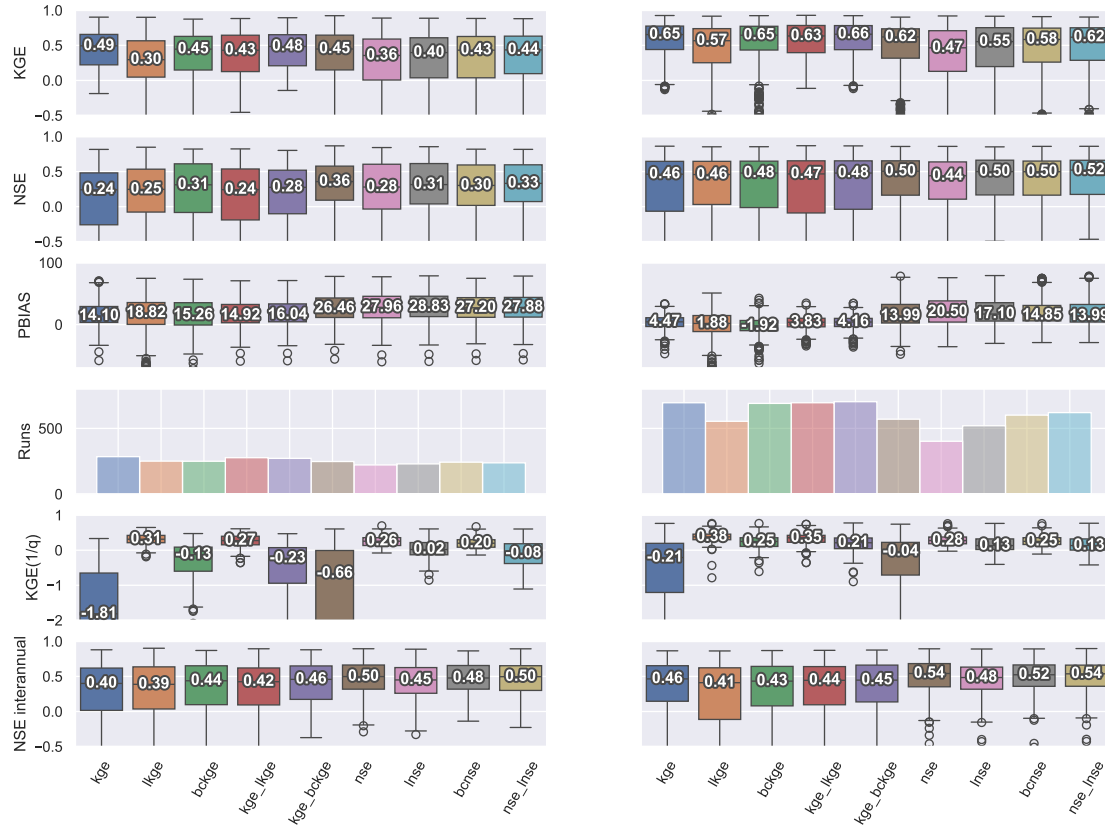


Figure 6. Boxplots for performance scores. Left panel: No precipitation correction, Right panel: with precipitation correction. From top to bottom: KGE, NSE, PBIAS, Number of Runs satisfying three criteria, KGE(1/q)

1.4 Hydrological regimes performance. Validation period only.

- 35 The Fig. 7 demonstrates that Atlantic regime has highest med(KGE) and lowest spread of the scores, whereas Mountain regime has the lowest med(KGE) and highest spread of the scores. Precipitation correction significantly improves med(KGE) for the Mountain regime.

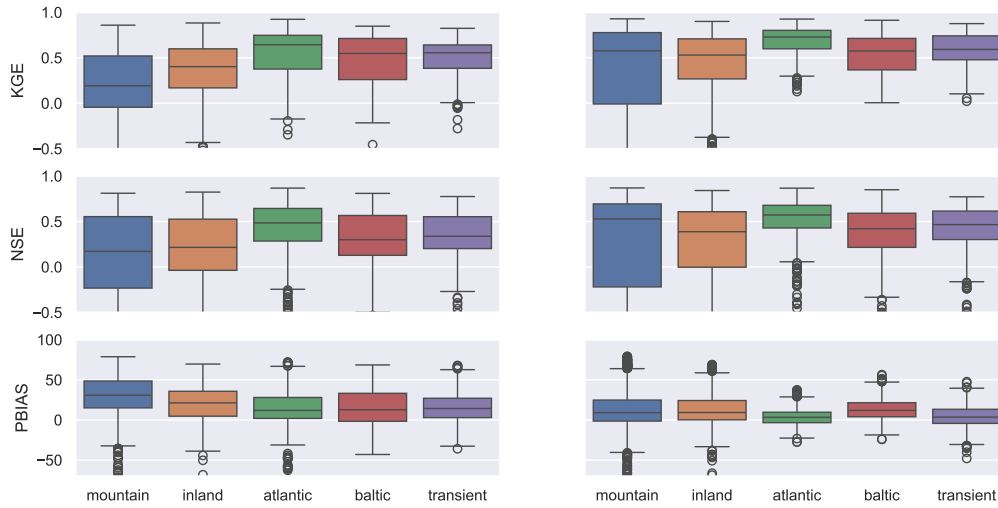


Figure 7. KGE performance scores per hydrological regime left: without precipitation correction, right: with precipitation correction

1.5 Heatmaps of model scores

- Figure 8 demonstrates heatmap of the NSE and KGE scores per catchment per goal function selection for **PTSTK** stack for the runs without precipitation correction. The target goal functions are split by vertical blue lines: KGE, LKGE, bcKGE, KGE_LKGE, KGE_bcKGE; the thick blue line split KGE-based target goal functions family from NSE-based family: NSE, LNSE, bcNSE, NSE_LNSE, NSE_bcNSE. As the stack is performing relatively well, it is difficult to find any connection between the performance and hydrological regime.
- 40

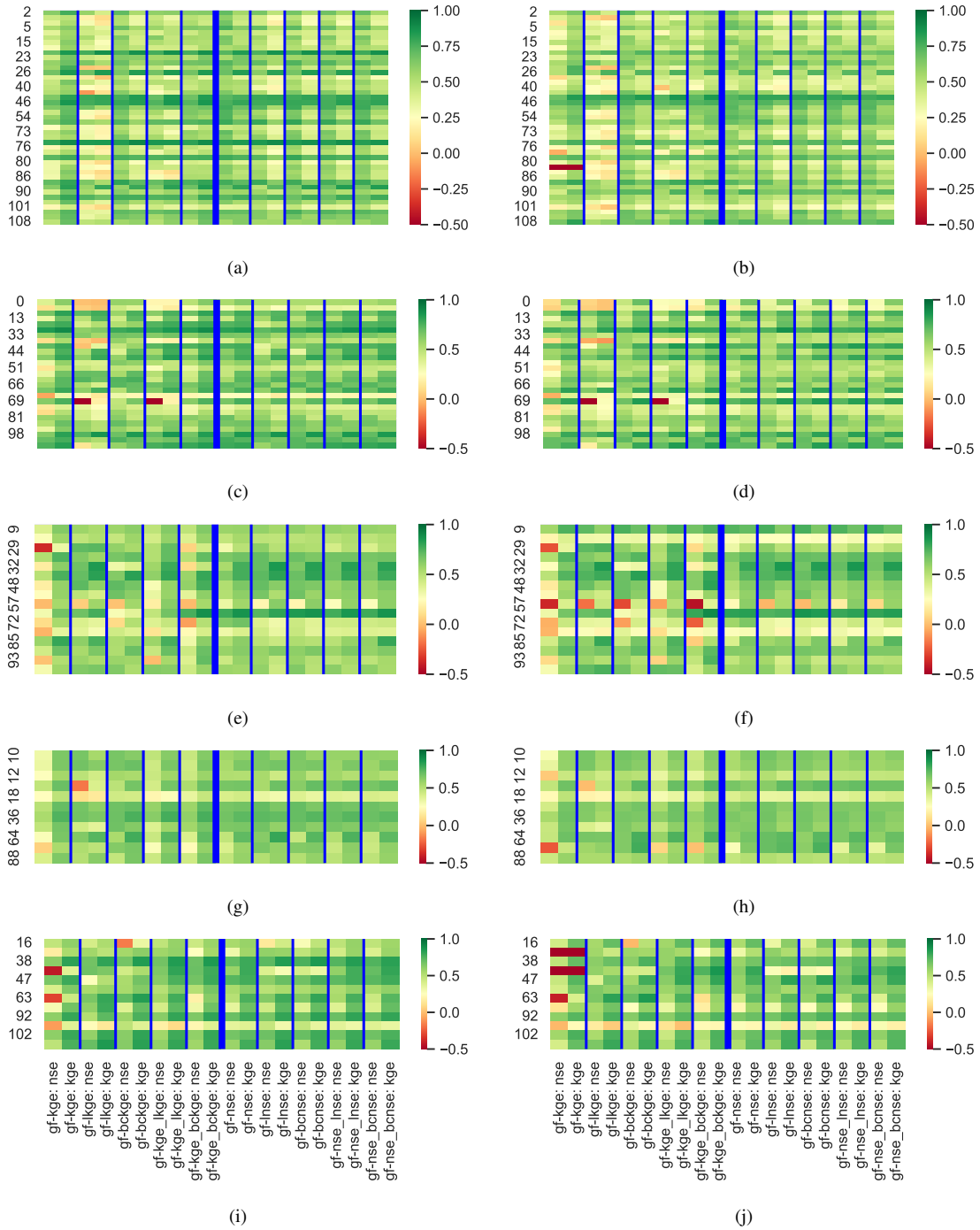


Figure 8. PTSTK stack heatmaps of KGE and NSE scores per hydrological regime, left column: calibration period, right column: simulations period, from top to bottom: mountain, inland, atlantic, baltic, transient hydrological regimes

For the **PTGSK** stack presented on fig. 9 the scores for catchment with the atlantic regime are noticeably higher than the
45 scores of the other regimes of the rest of the catchment population, followed by catchments with baltic and transient regimes.
The hydrological regimes of the poorest performing catchments are mountain and inland, with mountain showing the lowest
scores for all the configurations of the target goal function.

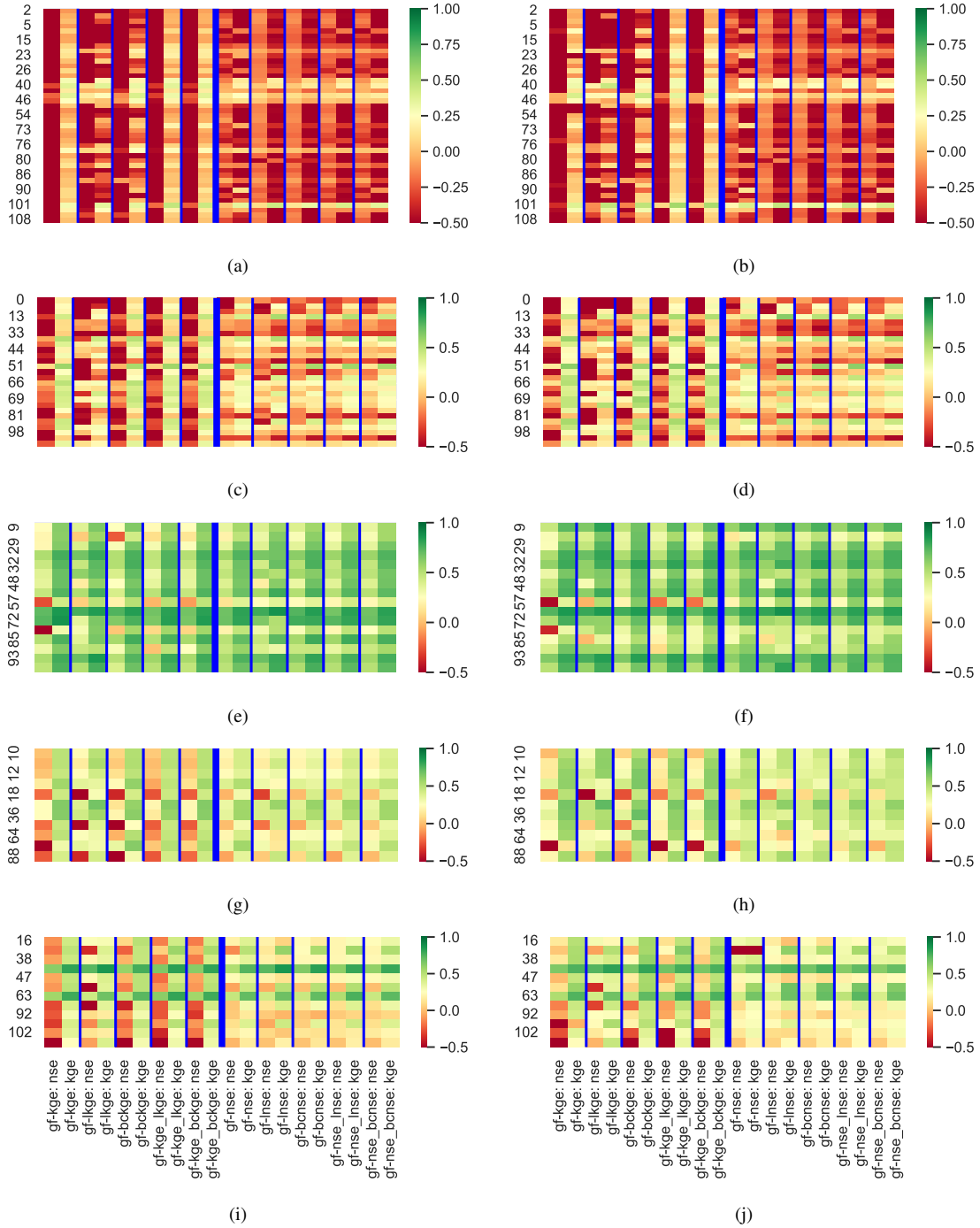


Figure 9. PTGSK stack heatmaps of KGE and NSE scores per hydrological regime, left column: calibration period, right column: simulations period, from top to bottom: mountain, inland, atlantic, baltic, transient hydrological regimes

1.6 Zero flow treatment

There has been a limited number of studies discussing zero flow treatment. As noted by Pushpalatha et al. (2012) adding a small value $\frac{1}{100}$ of mean streamflow makes it possible to compute some transformations (like logarithmic), when simulated or observed flows are zero. Discarding zero-flows and considering them as gaps is another approach. During parameter estimation for GR4J model (Westra et al., 2014) used threshold of 0.09 mm for streamflow for the catchment of appr.29 km² to avoid zero-flow issues. Another work suggests a need for special care of no flow conditions at intermittent and ephemeral catchments in Australia, where high proportion of streamflow is zero, for probabilistic hydrologic models (McInerney et al., 2019).

The definition of zero flow largely depends on the size of the catchment and quality of instrumentation used to measure streamflow. For the small catchments the value might be as small as 0.01 m³/s, whereas for medium and large catchments the zero flow might be anything between 0.1 and 1 m³/s. Everything below defined zero flow should be filtered out, so there is no extra numerical noise. But, if too much data is filtered out, the calibration might not be able to properly fit low flow, as well. For example, for catchment **8.6.0**, which has area of about 6 km² threshold of 0.01 leads to 12% of discharge data filtered out. In this study there are catchments with mean area of less than 3 km² and mean streamflow of app. 0.1 m³/s.

In addition, zero or no flow definition is important when specifying target discharge and goal function for calibration procedure. Transformations like logarithmic can create too much noise from near-zero periods of the flow, making it impossible for optimizer to fit the observations. However, for large scale studies it is difficult to define zero flow suitable for all catchments. This contributes to uncertainties in simulation of low flow.

The quality control and filtering of observations in Shyft happens before the data is put into the distributed time series service (DTSS) responsible for feeding computations with data. Thus, we performed a zero-flow experiment, before the main set of experiment.

Figure 10a) demonstrates heatmap of performance metrics for **PTSTK** model configuration during calibration for observed discharge filtered with threshold of 0.0001. Each 2 columns of heatmap corresponds to NSE and KGE performance metrics, where numbers 1-6 correspond to selection of target goal function during optimisation: 1. KGE, 2. LKGE, 3. KGE_LKGE, 4. NSE, 5. LNSE, 6. NSE_LNSE. The thick blue line splits KGE-based experiments from NSE-based experiments. Figure 10b) demonstrates heatmap of performance metrics for **PTSTK** model configuration during calibration for observed discharge filtered with threshold of 0.1. As expected the KGE based metric is most sensitive to the zero flow definition, especially experiments with target goal function KGE calculated on log-transformed flow (NSE2 and KGE2 metrics). As can be seen on the figure the simulation with higher threshold led to deterioration of performance for some of the catchments.

Based on this initial test we decided to use simple censoring approach. For further analysis the selection of catchments were based on availability of discharge observations, where 5% missing values were allowed, with a threshold 0.0. This allows us to put all the available discharge observations to DTSS. Our final selection of threshold ($q > 0.0$) is subjective and might not be optimal for all of the catchments, especially when log-transformed KGE as a goal function is selected, as it can create overfitting problems or possible non convergence issues in the optimization process: when the flow is near zero the KGE calculated on transformed flows generate very high negative values. In addition, it does not take into account possible lower/upper detection

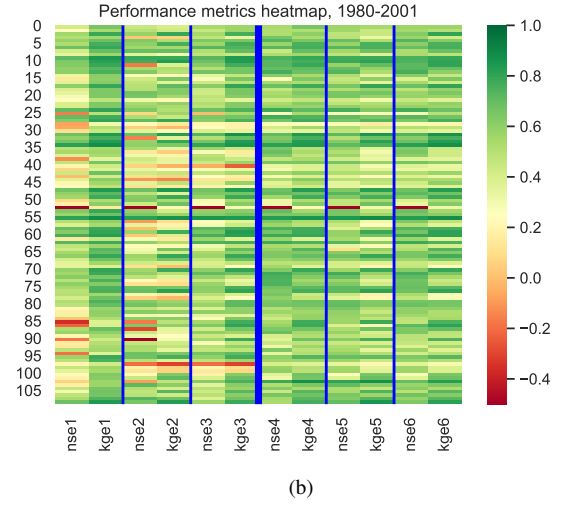
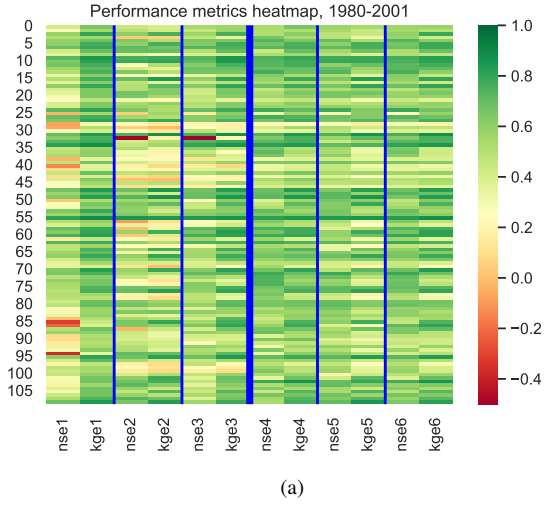


Figure 10. Results of simulation for PTSTK stack with different options of observed discharge filtering. Left to right: $q > 0.0001$, $q > 0.1$, 1-6 numbers correspond to goal function selection: 1. KGE, 2. LKGE, 3. KGE_LKGE, 4. NSE, 5. LNSE, 6. NSE_LNSE.

limits of instrumentation, when the measurement become unreliable. This is a subject for further studies. However, it is a set up, which is suitable for the most of the catchment population in the study.

References

- 85 Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, 377, 80–91, 2009.
- Knoben, W. J. M.: Setting expectations for hydrologic model performance with an ensemble of simple benchmarks, *Hydrological Processes*, 38, <https://doi.org/Setting expectations for hydrologic model performance with an ensemble of simple benchmarks>, 2024.
- Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments, *Water Resources Research*, 56, <https://doi.org/10.1029/2019WR025975>, 2020.
- 90 McInerney, D., Kavetski, D., Thyer, M., Lerat, J., and Kuczera, G.: Benefits of Explicit Treatment of Zero Flows in Probabilistic Hydrological Modeling of Ephemeral Catchments, *Water Resources Research*, 55, <https://doi.org/10.1029/2018WR024148>, 2019.
- Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, *Journal of Hydrology*, 11, <https://doi.org/10.1016/j.jhydrol.2011.11.055>, 2012.
- 95 Ruzzante, S. W., Knoben, W. J. M., Wagener, T., Gleeson, T., and Schnorbus, M.: Technical Note: High Nash Sutcliffe Efficiencies conceal poor simulations of interannual variance in tropical, alpine, and polar catchments, *egusphere*, <https://doi.org/10.5194/egusphere-2025-3851>, 2025.
- Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, *Hydrology and Earth System Sciences*, 22, <https://doi.org/10.5194/hess-22-4583-2018>, 2018.
- 100 Thirel, G., Santos, L., Delaigue, O., and Perrin, C.: On the use of streamflow transformations for hydrological model calibration, *egusphere*, preprint, <https://doi.org/10.5194/egusphere-2023-775>, 2023.
- Westra, S., Thyer, M., Leonard, M., Kavetski, D., and Lambert, M.: A strategy for diagnosing and interpreting hydrological model nonstationarity, *Water Resources Research*, 50, <https://doi.org/10.1002/2013WR014719>, 2014.