hess-2022-380: CC2

Parameter optimization of a hydrologic model needs to specify an objective or penalty function for the model to meet.

The classical Nash and Sutcliffe 1970 efficiency scale (NSE) expressed by Eq. (10) can be recast using the original notation as: $R^2 = 1 - F/F_0$. It has both an objective function in residual variance F, which is sum of squares of the simulation error (SSE) and an observed-mean-flow (μ_0) benchmark embedded in initial variance F_0, a fixed value. There is a one-to-one correspondence between NSE and F, and optimizing NSE is same as optimizing F. But this is not necessarily true in its variants, including an earliest known one, Ding 1974, Eqs. (40) and (47) therein.

NSE is a measure of correlation as well as others between simulation and observation as shown in a componentized form in Eq. (11). What it needs physically as well as statistically is at least one auxiliary benchmark to help interpret its intermediate scores between a perfect score of 1 for an observed or reference hydrograph, i.e. a perfect model, and of 0 for the (primary) benchmark model, μ_0 Establishing auxiliary benchmarks or baselines will help address one question about the popular performance metric: how close to 1 are NSE values reachable by models, e.g., Nearing et al. 2022, Table 1 therein.

The concept of two-parameter (ω_1:$\omega 2$) homothetic transformation hydrographs represents a first step toward searching for such auxiliary benchmarks, as described for a twin-peak synthetic hydrograph in Sections 3.1, 3.2, Equation (21), and presented in Figures 1, 2 and 3.

I've put forward a simplest second-order autogressive process of the streamflow, AR(2, c = 0, c1 = 2, c2 = −1), as a replacement of the primary benchmark, μ_0, e.g., Ding 2018. This, a slope-based projection hydrograph, instead could be considered a secondary benchmark, e.g., Azmi et al. 2021, SC1 and AC1 therein. In the same vein, a simplest third-order AR(3, 0, 2, −2, 1), a curvature-based projection hydrograph, could be a tertiary one.

AR(2) and AR(3) projection hydrographs can be generated for the twin-peak example hydrograph. Scoring them would yield NSE values, calibration free.

I encourage the authors to pursue this AR projection approach in a future study. For the example hydrograph, I for one would be interested in what are NSE scores for AR(2) and AR(3) benchmarks, and whether the higher score of the two is lower than but close to the values shown in Fig. 3(a) for both BB (Bad-Bad) and BG(Bad-Good) transformations.

Thank you for taking the time to review our manuscript and providing interesting explanations and suggestions about the use of autoregressive projections as a benchmark model.

We also read your comments in the peer-review section of Mizukami et al. (2019), Knoben et al. (2019), and Azmi et al. (2021), which were really insightful into the autoregressive projections. In a future study, it could be interesting to look for this aspect of performance criteria.

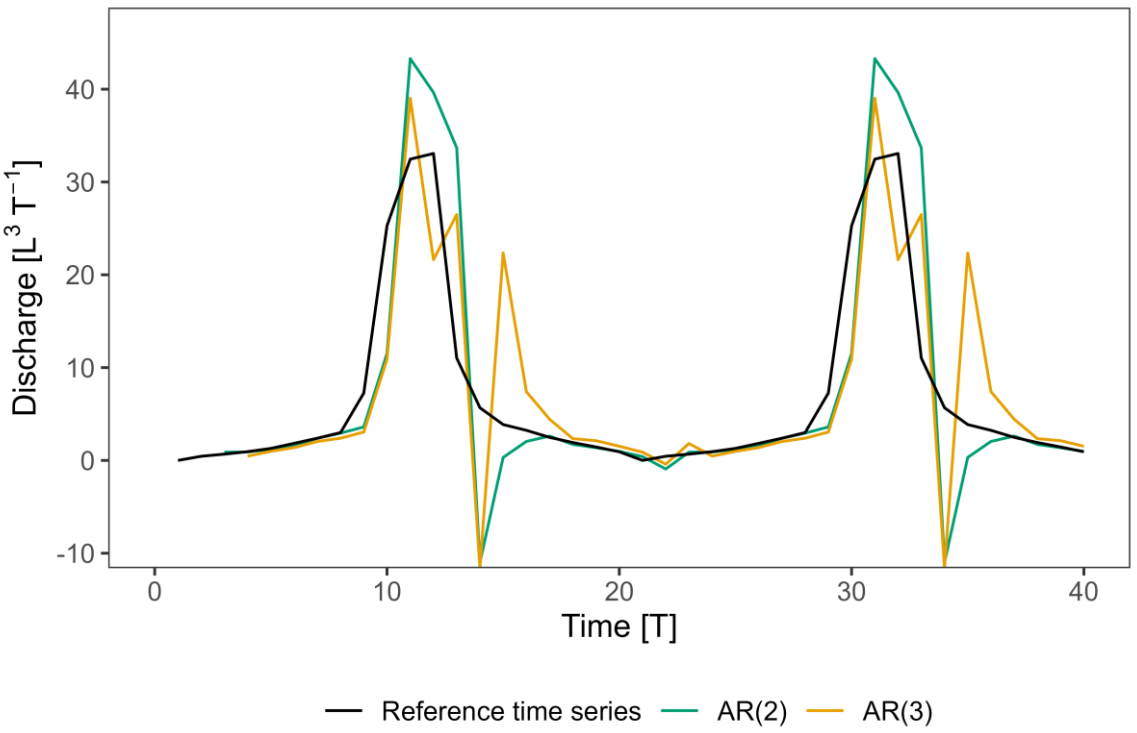
As you are interested in the results of the NSE scores for AR(2) and AR(3) benchmarks, we performed the calculation on the twin-peak example hydrograph, using the equations below:

$$Q_{AR2}(t) = 2 * Q_{obs}(t-1) - Q_{obs}(t-2)$$

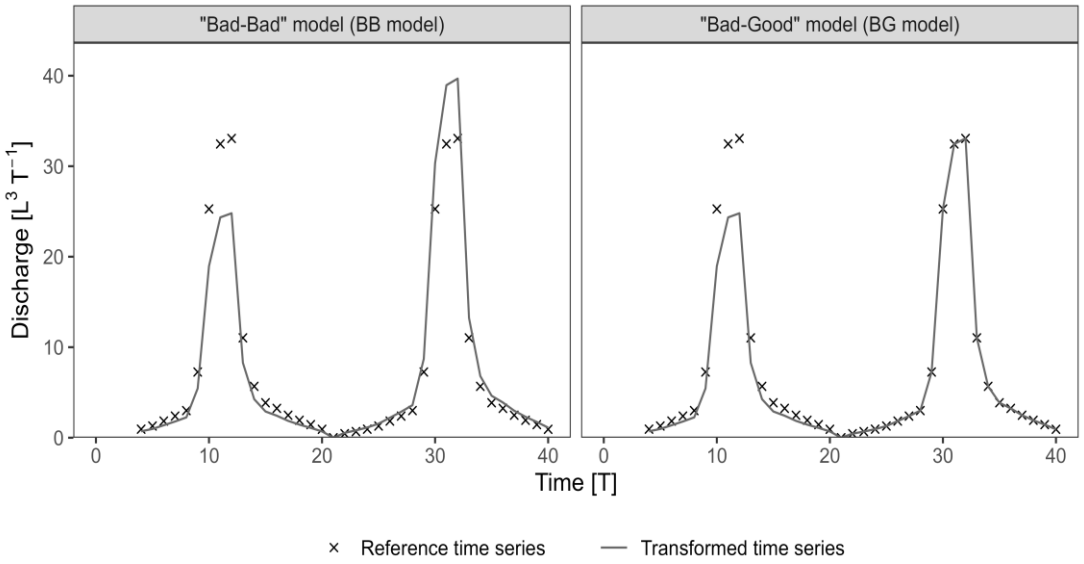
$$Q_{AR3}(t) = 2 * Q_{obs}(t-1) - 2 * Q_{obs}(t-2) + Q_{obs}(t-3)$$

$$NSE_{AR} = 1 - \frac{\sum\left(x_s(t) - x_o(t)\right)^2}{\sum\left(x_o(t) - Q_{AR}(t)\right)^2}$$

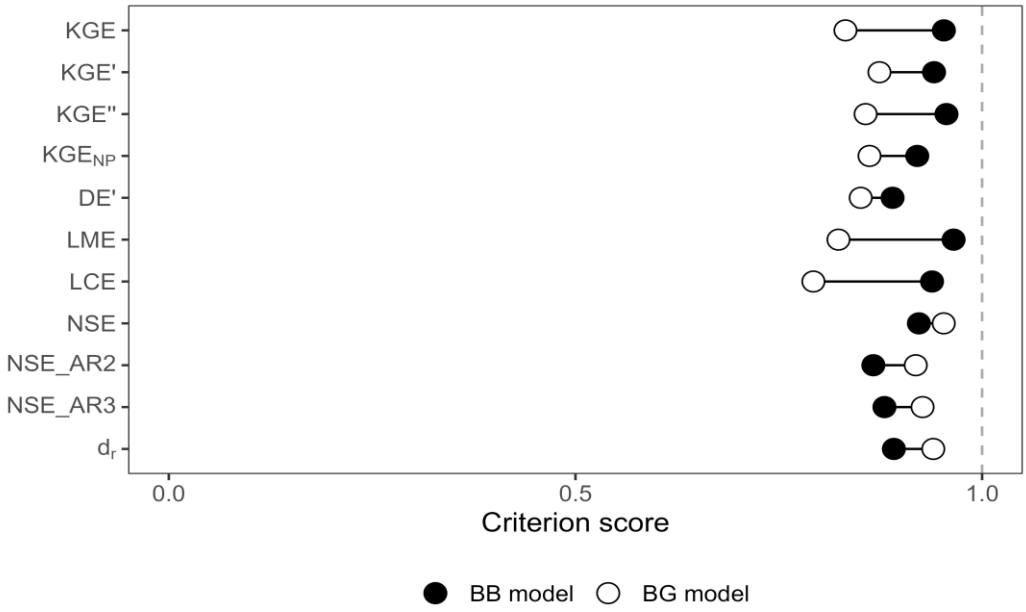
The following graph shows the observed time series alongside the time series of the different benchmarks used, i.e. $AR(2)$ and $AR(3)$:



Note that, the synthetic time series of the example hydrograph has no value before $t = 0$, therefore the first three values of the autoregressive projections will be non-defined as $Q_{obs}(t-1)$, $Q_{obs}(t-2)$ and $Q_{obs}(t-3)$ are *NA*. Because of this, we evaluated the models on the whole synthetic time series except the first three values, as can be appreciated on the graph below:



The following graph shows the score of each *NSE*, NSE_{AR2} and NSE_{AR3} performance criteria on the BB and BG synthetic models:



The values of the NSE, NSE_AR2 and NSE_AR3 are detailed in the table below:

Criterion	Bad-Bad model	Bad-Good model
NSE	0.922	0.953
NSE_AR2	0.866	0.918
NSE_AR3	0.880	0.927

We can see that the NSE with AR(3) benchmark has a higher score than with AR(2), with 0.880 and 0.927 for the BB and BG models, respectively. The NSE evaluations with AR benchmarks still yield good scores, close to the NSE with μ_0 benchmark. The score difference between the BB and BG models is slightly, but not significantly, higher for the NSE scores with AR benchmarks.

References

Azmi, E., Ehret, U., Weijs, S.V., Ruddell, B.L., Perdigão, R.A.P., 2021. Technical note: "Bit by bit": A practical and general approach for evaluating model computational complexity vs. Model performance. Hydrology and Earth System Sciences 25, 1103–1115. https://doi.org/10.5194/hess-25-1103-2021

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash and Kling efficiency scores. Hydrol. Earth Syst. Sci. 23, 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Mizukami, N., Rakovec, O., Newman, A.J., Clark, M.P., Wood, A.W., Gupta, H.V., Kumar, R., 2019. On the choice of calibration metrics for "high-flow" estimation using hydrologic models. Hydrol. Earth Syst. Sci. 23, 2601–2614. https://doi.org/10.5194/hess-23-2601-2019