A better metric than the NSE (Nash-Sutcliffe efficiency) ?

Heudorfer et al. (2026) achieve a higher median-performance of NSE=0.718 than the CAMELS one of NSE=0.696, by including better static-features data of their catchments (Table 2; Abstract, Lines 25 - 27). But better LSTM architecture raises the NSE by only 0.003 to 0.721 (Table 3).

To raise the calibrated NSE values to a higher median one, say 0.73, one may need to think outside the two boxes of static-features data and LSTM-based architecture.

The NSE, a model performance measurement standard, is well-known for its iconic benchmark, the mean of observed discharges. In the context of the Nash-Sutcliffe efficiency criterion, to develop a better scale is to replace the benchmark with a physically realistic one on riverine watersheds.

To help guide a model's calibration, this writer previously proposed an alternate benchmark, called the AR2, to the Nash-Sutcliffe one, Cinkus et al. (2023: CC2 by Ding; AC2). The AR2 benchmark is represented by a streamflow projection function as follows:

$$Q_{AR2}(t+1) = 2Q_{obs}(t) - Q_{obs}(t-1)\,.$$

Q_{AR2}, the AR2 benchmark, is an observations-driven one. (The subscript AR2 stands for a nominal 2nd-order, AutoRegression of the discharges.)

Replacing the Nash-Sutcliffe benchmark by the AR2 one, the mean-based NSE scale becomes an acceleration-based AR2 one. The benchmark substitution represents a paradigm shift from a static measure, the NSE scale, to a dynamic one, the AR2.

The two scales are related by equating the common SSE (sum of squares of error) term,

whose minimization, i.e. the 'Least Squares' solution, is a first objective of a model's calibration.

From a synthetic twin-peak hydrograph (Good-Good model), Cinkus et al. (2023) formed two models varying in peak magnitude: a lower-higher (i.e. Bad-Bad) model and a lower-same (i.e. Bad-Good) model. Their calibrated NSE and AR2 values are shown below:

Scale	Bad-Bad model	Bad-Good model	Good-Good model
NSE	0.922	0.953	1.0
AR2	0.866	0.918	1.0

In comparison with the NSE scale, the AR2 one scores lower values for two variant models, but has a wider range. Because the acceleration-based AR2 scale is more sensitive to change in a

model's configuration or architecture, this appears a better scale than the static NSE one.

Whether or not an LSTM can match the AR2 benchmark is an open question, which has been addressed tentatively by Cinkus et al. (2023, unpublished data).

References

Cinkus, G., Mazzilli, N., Jourde, H., Wunsch, A., Liesch, T., Ravbar, N., Chen, Z., and Goldscheider, N.: When best is the enemy of good – critical evaluation of performance criteria in hydrological models, Hydrol. Earth Syst. Sci., 27, 2397–2411, https://doi.org/10.5194/hess-27-2397-2023, 2023.

Heudorfer, B., Gupta, H., Dolich, A., and Loritz, R.: Better data or better architecture? Improving deep-learning-based prediction in ungauged basins, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2026-1965, 2026.