Review of HESS Technical note: "*Inherent benchmark or not? Comparing Nash- Sutcliffe and Kling-Gupta efficiency scores*", by Wouter J, M Knoben, JE Freer and RA Woods

Review Provided by Hoshin Gupta (23rd July 2019)

Summary of the Paper: The paper makes perhaps three main points:

Main Point Number (1): On Use of the "Mean Flow Benchmark" to interpret NSE and KGE

- The NSE normalizes model performance to an interpretable scale such that NSE = 1 indicates perfect correspondence between simulations and observations, NSE = 0 indicates that the model simulations have the same explanatory power as the mean of the observations, and NSE < 0 indicates that the model is a worse predictor than the mean of the observations.
- NSE = 0 is regularly used as a benchmark to distinguish '*good*' and '*bad*' models, although this threshold could be considered a low level of predictive skill and is also a relatively arbitrary choice.
- KGE addresses several shortcomings in NSE and is increasingly used for model calibration and evaluation. Like NSE, KGE = 1 indicates perfect agreement between simulations and observations.
- Some users have tried to assign a similar scale/threshold as with NSE to be used in interpretation of KGE scores. Many authors use positive KGE values as indicative of '*good*' model performance, and negative KGE values as indicative of '*bad*' performance.
- However, this paper shows that placing the threshold for '*good*' model performance at KGE = 0 is generally correct (i.e., positive KGE values do indicate improvements upon the mean flow benchmark) but not complete. In fact, *negative KGE values do not necessarily indicate a model that performs worse than the mean flow benchmark*. The authors show this in mathematical terms, and then present results from a synthetic experiment to highlight that NSE and KGE values are not directly comparable and that understanding of the NSE metric does not translate well into understanding of the KGE metric.
- Mathematically, if the model simulations of the system responses are in fact constant over time and equal to the mean of the observed flows (the mean flow benchmark), we actually have KGE ≈ −0.41.

Main Point Number (2): On the Need to Explicitly Consider Benchmark Performance

- NSE and KGE values cannot be directly compared and should not be treated as approximately equivalent. There is no unique relationship between NSE and KGE values and where NSE values fall in the KGE component space depends in part on the coefficient of variation (CV) of the observations.
- NSE values that are traditionally seen as high do not necessarily translate into high KGE values. Hydrologists who choose to use the KGE metric need to develop new understanding of how this metric should be interpreted and not let themselves be guided by their understanding of NSE.

- Whereas NSE has an inherent benchmark in the form of the mean flow, this benchmark is not inherent in the definition of KGE, which is instead an expression of distance away from the point of ideal model performance in the space described by its three components.
- There is no direct reason to use the "*mean flow*" as a benchmark over other options.
- Because KGE has no inherent benchmark value to enable a distinction between '*good*' and '*bad*' models, *modelers using KGE must be explicit about the benchmark model or value they use to compare the performance of their model against*.
- By choosing the mean flow as a benchmark to distinguish between '*good*' and '*bad*' models, practitioners limit themselves in the models and/or parameter sets they consider in a given study, without rational justification.

Main Point Number (3): On the Need to Recognize that Metrics and Benchmarks are Purpose-Dependent

- There is no single perfect model performance metric that is suitable for every study purpose. Indeed, global metrics that lump complex model behaviour and residual errors into a single value are not useful for exploring model deficiencies and diagnostics regarding how models fail or lack certain processes.
- In the choice of metrics, modellers should make conscious and well-founded choices about which aspects of the simulation they consider most important (if any), and in which aspects of the simulation they are willing to accept larger errors.
- When using KGE, emphasizing certain aspects of a simulation is straightforward by attaching weights to the individual KGE components to reduce or increase the impact of certain errors on the overall KGE score.
- This purpose-dependent score should then be compared against a purpose-dependent benchmark to determine whether the model can be considered '*good*'.
- How these purpose-dependent benchmarks should be set is an open question to the hydrologic community.

My Review Remarks:

[1] I thoroughly enjoyed reading this Technical Note contribution by *Wouter, Knoben, Freer* and *Woods*, and I thank them for (re)raising some very important issues, and for their new/original contribution regarding the value that the KGE criterion takes on when using the mean flow as a benchmark. As such, I have no critique per se to offer regarding this paper, and compliment the authors on an excellent contribution to the literature.

[2] Instead I would like to focus on some interesting points raised by this work. This review opportunity allows me to take the liberty of reminding the readers of some interesting points that were previously raised in *Schaefli and Gupta (2007)* and *Gupta et al (2009)*, that the authors

allude to, *but which perhaps could be strengthened by the authors of the current work in their presentation*. Text between quotes is reproduced from the original papers.

[3] Beginning first with *Schaefli and Gupta (2007)*, that paper was about benchmarking. In it, we discussed the fact that the process by which anyone assesses and communicates model performance evaluation is of primary importance, and that "*the basic 'rule' is that every modelling result should be put into context, for example, by indicating the model performance using appropriate indicators, and by highlighting potential sources of uncertainty*".

[4] We pointed out therein (as have others before and after us) that:

a) the *"NSE value, while a convenient and normalized measure of model performance does not provide a reliable basis for comparing the results of different case studies*"
b) the "*use of the mean observed value as a reference can be a very poor predictor (e.g. for strongly seasonal time series), or a relatively good predictor (e.g. for time series that are essentially fluctuations around a relatively constant mean value)*".

For example, "*In the case of strongly seasonal time series, a model that only explains the seasonality but fails to reproduce any smaller time scale fluctuations will report a good NSE value; for predictions at the daily time step, this (high) value will be misleading. In contrast, if the model is intended to simulate the fluctuations around a relatively constant mean value, it can only achieve high NSE values if it explains the small time-scale fluctuations*".

[5] Therefore, the definition of an appropriate benchmark model is particularly important … to properly communicate how good a model really is, it is necessary to establish an appropriate reference (or benchmark model) for a given case study and a given modelling time step. In that paper we mention some examples, including:

a) the interannual mean value for every calendar day proposed by *Garrick et al. (1978)* for systems having strong but relatively constant seasonality
b) a simple adjusted precipitation benchmark (APB) where the rainfall is scaled to match the mean discharge and shifted in time by some optimum lag that reflects the time of concentration of the basin, and
c) a smoothed version of the APB where a simple dispersion process (moving average) is added to adjust the smoothness of the scaled-down and translated precipitation to match the smoothness of the observed discharge, for example by maximizing the correlation between the adjusted precipitation and the observed flow *(Morin et al., 2002)*.

Of course, many other possible benchmarks can be conceived, such as "*persistence*" (the next time steps' simulated flow is the same as the current time step's observed flow), some kind of linear or non-linear extrapolation into the future, and some kind of data-based time-series analytical model projection such as can be constructed by ARMAX or ANN methods.

[6] In the conclusions to *Schaefli and Gupta (2007)*, we argued that the definition of an appropriate baseline for model performance, and in particular, for measures such as NSE (and by extension, KGE or any other model performance measure), should become part of the 'best practices' in hydrologic modelling, that "*Every modelling study should explain and justify the choice of benchmark*", and that "*the benchmark should fulfill the basic requirement that every*

hydrologist can immediately understand its explanatory power for the given case study and, therefore, appreciate how much better the actual hydrologic model is".

[7] Moving next to *Gupta et al (2009)*, we discussed the fact that the NSE, which is a dimensionless mathematical normalization of the mean squared error (MSE) criterion can be viewed as a classic skill score (*Murphy, 1988*), where 'skill' is interpreted as the comparative ability of a model with regards to a baseline 'model'. Further, as shown by *Murphy (1988)* and *Weglarczyk (1998)*, it is possible to decompose the NSE criterion into components (correlation, conditional bias, and unconditional bias) that facilitates a better understanding of what is causing a particular model performance to be 'good' or 'bad', while providing insight into possible trade-offs between the different components.

[8] Our own particular diagnostic decomposition of NSE (and hence MSE) was developed in the context of our interest in hydrological modelling where, as we showed, interactions among these components (correlation, mean bias, and variance bias) can cause problems during model calibration – possibly leading to parameter estimates that are associated with large volume balance errors and/or underestimation of the variability in the flows. Further, we pointed out that many different combinations of the three components can result in the same overall value for NSE, leading to considerable ambiguity in the comparative evaluation of alternative model hypotheses.

[9] Importantly, we also pointed out that, rather than trying to come up with a 'corrected' version of the NSE criterion, *the whole calibration problem can instead be viewed from the multi-objective perspective* (see e.g., *Gupta et al., 1998*), by focusing on the correlation, variability error and bias error as separate criteria to be optimized. *When we do so, if a compromise solution is desired, we can use the solution provided by the KGE or one of its alternatively weighted variants*.

[10] We presented some comparative experimental results that show that when optimizing on KGE, there is a strong correlation between the values obtained for the KGE and NSE criteria, but when optimizing on NSE, the correlation between the values obtained for NSE and KGE is lower due to the fact that optimization on KGE strongly controls the values that the mean and variance ratio components can achieve, whereas optimization on NSE constrains these components only weakly. Overall, the use of KGE instead of NSE for model calibration tends to improve the bias and variability measures considerably while only slightly decreasing the correlation.

[11] Finally, we pointed out that the NSE/MSE or KGE performance metric decomposition relates to the idea of diagnostic model evaluation, as proposed by *Gupta et al. (2008)*, *which is to move beyond aggregate measures of model performance that are primarily statistical in meaning, towards the use of (multiple) measures and signature plots that are selected for their ability to provide hydrological interpretation*. While the theoretical development behind the KGE provides one simple, statistically founded approach to the development of a strategy for diagnostic evaluation and calibration of a model, *we also pointed out that all other statistical properties beyond the mean and standard deviation (which are two long-term statistics of the data), such as timing of the peaks, and shapes of the rising limbs and the recessions of the hydrograph (i.e. autocorrelation structures), are lumped into the (linear) correlation coefficient as an aggregate measure*.

[12] We therefore suggested that a logical next step would be to consider other relevant diagnostic properties (such as for example, different aspects of flow timing and shape), but left those considerations are left for future work. For example, although not mentioned explicitly in *Gupta et al (2009)*, there is no reason that other (statistical or otherwise) aspects of model performance, such as "*skewness*", or "*particular quantiles*" etc., should not be integrated into the basis for model performance evaluation and, if desired, built into a "*KGE-like*" metric.

[13] However, the *explicitly stated purpose* of the *Gupta et al (2009)* study *was NOT to design an improved measure of model performance*, but instead:

a) to show clearly that there are systematic problems inherent with any optimization that is based on mean squared errors (such as NSE),
b) that "*the alternative criterion KGE was simply used for illustration purposes*" (many different alternative criteria would also be sensible), and
c) that "*Ultimately the decision to accept or reject a model must be made by an expert hydrologist, where such a decision is best based in a multiple-criteria framework*", where tracking the mean bias, variance bias and correlation (and other possible) components can help.

Concluding Remarks:

[14] With this context, it would actually be useful for the community to strategically move beyond the use of single metrics for model performance assessment (and/or selection), whether NSE or KGE or any other that might be conceived, and to follow the spirit of *Gupta et al (2008)* by designing some reasonable and rational basis for selecting "*sets*" of metrics that provide meaningful diagnostic evaluation of a model.

[15] As pointed out by the current authors, to be meaningful, any such metrics should be accompanied by meaningful benchmarks. To be meaningful, these benchmarks should *not* be specified in an ad-hoc manner (such as NSE > 0.5 etc.) but should have some meaningful theoretical basis that conveys useful information to the decision maker.

[16] Indeed, I have often been contacted by researchers asking for some "*threshold*" values to use with KGE in their studies, and have always responded by discouraging such a practice and instead encouraging the use of the individual diagnostic components of KGE (and others that might be imagined) and setting associated thresholds using some meaningful basis.

[17] I do understand that, when performing studies involving large samples of data and/or many models, there is a tendency to want to use simple "*aggregate*" metrics in order to select or focus on a sub-set of "*good*" or "*poor*" models. However, there is arguably little to be gained by doing so by following the (arguably lazy) approach of using an aggregate metric that is not meaningfully interpretable.

[18] I sincerely hope that this current authored contribution will help to move the bulk of the community of hydrologic practitioners in the direction of using a more informative, and powerful, *diagnostic* (and necessarily multi-criteria) basis for model evaluation *that points to the nature of model deficiencies* and therefore to the modeling issues that need fixing.

[19] It might be helpful therefore, for the current authors to make some stronger arguments/comments in this direction, to encourage movement beyond the use of NSE and/or KGE, and thereby to a more powerful and robust approach to model assessment, as has been (slowly) pursued the case in some closely related communities (*Abramowitz 2012*).

References Cited:

Abramowitz G, 2012, Towards a public, standardized, diagnostic benchmarking system for land surface models, Geoscientific Model Development, vol. 5, pp. 819 - 827, http://dx.doi.org/10.5194/gmd-5-819-2012

Garrick M, Cumane C, Nash JE. 1978. A criterion of efficiency for rainfall-runoff models. Journal of Hydrology 36: 375–381.

Gupta HV, S Sorooshian and PO Yapo (1998), Towards Improved Calibration of Hydrologic Models: Multiple and Non-Commensurable Measures of Information, Water Resources Research, Vol. 34, No. 4, pp. 751-763

Gupta HV, T Wagener and YQ Liu (2008), Reconciling Theory with Observations: Towards a Diagnostic Approach to Model Evaluation, Hydrological Processes, Vol. 22 (18), pp. 3802-3813, DOI: 10.1002/hyp.6989.

Gupta HV, H Kling, KK Yilmaz and GF Martinez-Baquero (2009), Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling, Journal of Hydrology, Vol. 377, pp. 80-91, doi: 10.1016/j.jhydrol.2009.08.003.

Morin E, Georgakakos KP, Shamir U, Garti R, Enzel Y. 2002. Objective, observations-based, automatic estimation of the catchment response timescale. Water Resources Research 38: 1212, DOI: 10·1029/2001WR000808.

Murphy A (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review 116: 2417-2424

Schaefli B and HV Gupta (2007), Do Nash values have value? Hydrological Processes, 21(15), 2075-2080, simultaneously published online as Invited Commentary in Hydrologic Processes (HP Today), Wiley InterScience, doi: 10.1002/hyp.6825

Weglarczyk S (1998), The interdependence and applicability of some statistical quality measures for hydrological models. Journal of Hydrology 206: 98-103