Supplement of Hydrol. Earth Syst. Sci., 28, 3665–3673, 2024
https://doi.org/10.5194/hess-28-3665-2024-supplement
© Author(s) 2024. CC BY 4.0 License.

Hydrology and Earth System Sciences
Open Access
EGU

Supplement of

Technical Note: The divide and measure nonconformity – how metrics can mislead when we evaluate on different data partitions

Daniel Klotz et al.

Correspondence to: Daniel Klotz (daniel.klotz@ufz.de)

The copyright of individual parts of the supplement might differ from the article licence.

S1 Exploring a situation-equitable Nash–Sutcliffe Efficiency

A modified NSE could also evaluate each sample differently and evaluate situations that are easy to predict more strictly and situations that are difficult to predict less strictly. We refer such a modification the Situation-Equitable NSE (SENSE). A specific implementation that uses the design principles from Sect. 2.2 of the original manuscript is

$$\mathrm{SENSE} = 1 - \frac{\sum_{t=1}^{T} \frac{1}{(o_t - \hat{\sigma}_t)^2}(o_t - s_t)^2}{\sum_{t=1}^{T}(o_t - \hat{\sigma}_t)^2}. \tag{S1}$$

Here, $\hat{\sigma}_t^2$ is an estimation for the observational variance at time t, which we estimate by using a nearest neighbor approach that draws from a reference set:

$$\hat{\sigma}_t = \frac{1}{K}\sum_{k=1}^{K} \mathrm{kNN}(\boldsymbol{c}_t, \mathrm{C_R}, k), \tag{S2}$$

where $\boldsymbol{c}_t = [o_{t-9}, o_{t-8} \ldots, o_{t-1}, o_t]$ is a situation vector containing the current observation and additional context in the form of preceding runoff values of the last 10 day, $\mathrm{bmC_R} = \{\boldsymbol{c}_k, \boldsymbol{c}_{k-1}, \ldots, \boldsymbol{c}_0\}$ is a reference storage, and kNN yields the last observation o_k within $\hat{\boldsymbol{c}}_k = [o_{k-9}, o_{k-8}, \ldots, o_{k-1}, o_k]$, which in itself is the k-th nearest neighbor of $\boldsymbol{c}_t$ within $\mathrm{bmC_R}$. That is, we try to weight each given timestep t by finding an approximation to the conditional variance of said timestep by taking the runoff observations in $\boldsymbol{c}_t$ and using the kNN regressor to find the k most similar runoff vectors. From these we then derive a situational estimation of the variance. Thus, the neighborhood of the kNN algorithm serves as the locality and its mean is our estimator for the expectation.

The SENSE would introduce two hyper-parameters: k and C and one could think about extend it to allowing arbitrary measures of similarity. Our analysis of the relative importance of choice k shows that SENSE that the parameter is not particularly sensitive (Fig. S1). If one wants to develop more sophisticated extensions one could use more nuanced similarity measures and include explanatory variables (e.g., meteorological forcings).

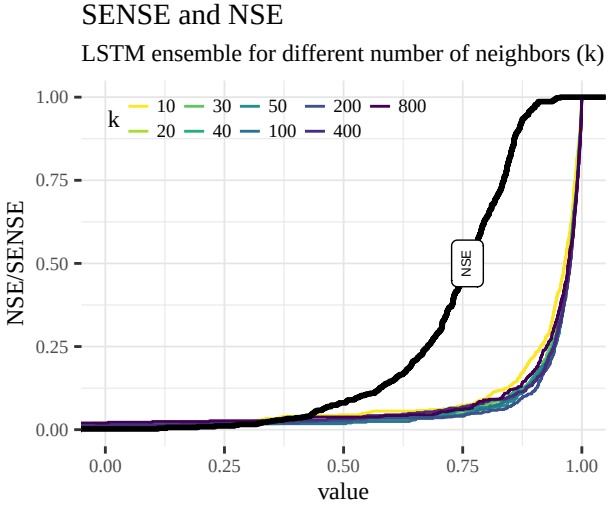

Figure S1. Approximation behavior of the SENSE, given the LSTM ensemble from Kratzert et al. (2019). The test data for a basin comprises 10 years of daily data (i.e., 3650 data points). Thus, 800 neighbors consider ca. 20% of the data for each time step.

S2 Connections between global and local performance criteria

In the following, we consider a dataset $D_T = \{(\boldsymbol{o}_t, \boldsymbol{s}_t) \mid t \in 1, 2, \ldots, T\}$ with tuples of observations $\boldsymbol{o}_t \in \mathbb{R}^m$ and simulation values $\boldsymbol{s}_t \in \mathbb{R}^m$ that correspond to the index t (which can, but not necessarily has to, correspond to a time step). In general $\boldsymbol{o}_t$ and $\boldsymbol{s}_t$ are vectors, but, for the sake of simplicity, in the following we will only discuss the special case where they are scalars. This choice does not result in any loss of generality.

Definition S2.1. A partitioning of D_T, with the number of partitions denoted by Z, is a sequence of disjoint sets $A_1, A_2, ...$ so that their union yields D_T. That is: $\bigcup_{z=1}^{Z} A_z = D_T$.

A specific partition A_z of D_T is hence given by $A_z = \{(o_t, s_t) \mid t \in 1, 2, \ldots, T \text{ and } 1_{(o_t, s_t) \in A_z} = 1\}$, where $1_{(o_t, s_t) \in A_z}$ is an indicator function that returns 1 if the criteria for the desired partition are met by the datapoint at t and 0 otherwise. We will always assume that the partitions are chosen so that a given performance criterion can be reasonably evaluated. To give an example: for the NSE this would mean that no partition with less than two data-points can be created (since then the sample variance of the observations would be undefined), and it is not allowed to choose the partitions so that all observation values are the same (again, because then the sample variance of the observation would be zero). For any practical application this is not a strong assumption (counterexamples do nevertheless exist, even in practise. For example, dry seasons in ephemeral streams and frozen streams could be special cases where partitioning can easily yield partitions with zero variance).

The indicator function for constructing A_z trivially implies $A_z \subseteq D_T$ and leads to a convenience index function $\mathbb{I}_{A_Z}$ that we will use to sum over the properties of a partition:

$$\mathbb{I}_{A_Z}(x) = \sum_{t=1}^{T} 1_{(o_t, s_t) \in A_z} * x_t, \tag{S3}$$

where x here is a placeholder variable. We also define the model error as $e_t = o_t - s_t$, and will, for convenience, omit the function arguments and brackets where it is clear from the context. For example:

1. $\mathbb{I}_{A_Z}(y) = \sum_{t=1}^{T} 1_{(o_t, s_t) \in A_z} * y_t$.
2. $\mathbb{I}_{A_Z}(1) = \sum_{t=1}^{T} 1_{(o_t, s_t) \in A_z}$ is the number of all elements in A_z (that is, the size of the set A_z).
3. $\mathbb{I}_{A_Z} e_t \coloneqq \mathbb{I}_{A_Z}(e_t) = \sum_{t=1}^{T} 1_{(o_t, s_t) \in A_z} * e_t$ is the sum of the errors (i.e., the bias) from the elements in A_z.

S2.1 Generality

Without loss of generality we will prove all the below properties with **two** partitions for the sake of simplicity. The results generalize to higher numbers of partitions by recursively applying the same logic.

To illustrate this, let L denote a to-be-minimized performance criterion (where we use L_{A_z} to express that we evaluate over the data in A_z), i.e. a loss, which obeys

$$L_{A_i} \leq L_{A_j} \Longrightarrow L_{D_T} \leq L_{A_j}, \tag{S4}$$

where A_i and A_j are two partitions of the dataset D_T. In words, the assumption from eq. S4 enforces that if the loss for one partition is smaller than the loss for a second one, then this implies that the loss for the overall data will also be smaller for the second partition.

Proposition S2.2. *This inequality also holds if we divide D_T into three partitions A_a, A_b, A_c. That is:*

$$L_{A_a} \leq L_{A_b} \leq L_{A_c} \implies L_{D_T} \leq L_{A_c} \tag{S5}$$

also holds.

Proof. To prove this, we use the fact that, by the assumption from eq. S4, if $L_{A_a} \leq L_{A_b}$ then this implies that $L_{A_a \cup A_b} \leq L_{A_b}$. Formally,

$$L_{A_a} \leq L_{A_b} \Longrightarrow L_{A_a \cup A_b} \leq L_{A_b},$$

and therefore

$$L_{A_a} \leq L_{A_b} \leq L_{A_c} \Longrightarrow L_{A_a \cup A_b} \leq L_{A_b} \leq L_{A_c}.$$

Further,

$$L_{A_a \cup A_b} \leq L_{A_c} \Longrightarrow L_{A_a \cup A_b \cup A_c} \leq L_{A_c}$$

also holds by the assumption from eq. S4, so we can conclude that Eq. (S5) is true. □

This nested form of evaluation can be repeated no matter how many partitions there are and it can also be repeated for different implications. Therefore, we can analyze the behavior of two partitions without loss of generality.

S2.2 Mean squared error

The sample MSE for a partition $A_X \in D_T$ is defined as:

$$\mathrm{MSE}(A_x) = \frac{\mathbb{I}_{A_X} e_t^2}{\mathbb{I}_{A_X}}. \tag{S6}$$

Proposition S2.3. *Given two partitions, $A_i \subset D_T$ and $A_j \subset D_T$, with $\mathrm{MSE}_{A_i} \leq \mathrm{MSE}_{A_j}$, the MSE_{D_T} is bound by $\mathrm{MSE}_{A_i} \leq \mathrm{MSE}_{D_T} \leq \mathrm{MSE}_{A_j}$.*

Proof. Expanding and rearranging $\mathrm{MSE}_{A_i} \leq \mathrm{MSE}_{A_j}$ gives:

$$\mathbb{I}_{A_J} * \mathbb{I}_{A_I} e_t^2 \leq \mathbb{I}_{A_I} * \mathbb{I}_{A_J} e_t^2.$$

If we expand on both sides by $\mathbb{I}_{A_I} * \mathbb{I}_{A_I} e_t^2$, we get:

$$\begin{aligned}
\mathbb{I}_{A_J} * \mathbb{I}_{A_I} e_t^2 + \mathbb{I}_{A_I} * \mathbb{I}_{A_I} e_t^2 &\leq \mathbb{I}_{A_I} * \mathbb{I}_{A_J} e_t^2 + \mathbb{I}_{A_I} * \mathbb{I}_{A_I} e_t^2, \\
(\mathbb{I}_{A_J} + \mathbb{I}_{A_I}) * \mathbb{I}_{A_I} e_t^2 &\leq \mathbb{I}_{A_I} * \mathbb{I}_{D_T} e_t^2, \\
\frac{\mathbb{I}_{A_I} e_t^2}{\mathbb{I}_{A_I}} &\leq \frac{\mathbb{I}_{D_T} e_t^2}{\mathbb{I}_{D_T}}, \\
\mathrm{MSE}_{A_i} &\leq \mathrm{MSE}_{D_T}.
\end{aligned}$$

Hence, the smaller MSE of the two partitions is also smaller than the MSE of the whole dataset. Inversely, the larger of the MSEs of the two partitions is larger than the MSE of the whole dataset — which can be shown analogously to the provided derivation. □

Thus, we can summarize the results of our proof with following relationship:

$$\mathrm{MSE}_{A_i} \leq \mathrm{MSE}_{D_T} \leq \mathrm{MSE}_{A_j}.$$

S2.3 Weighted mean squared error

The sample weighted mean squared error WMSE for a partition $A_X \in D_T$ is defined as:

$$\mathrm{WMSE}(A_x) = \frac{\mathbb{I}_{A_X} w_t * e_t^2}{\mathbb{I}_{A_X} w_t}, \tag{S7}$$

where w_t are the weights given to each individual sample.

Proof. The proof is analogous to proof S2.2. Despite the redundancy we show it in the following for the sake of completeness.

Expanding and rearranging $\mathrm{WMSE}_{A_i} \leq \mathrm{WMSE}_{A_j}$ gives:

$$\mathbb{I}_{A_J} w_t * \mathbb{I}_{A_I} w_t e_t^2 \leq \mathbb{I}_{A_I} w_t * \mathbb{I}_{A_J} w_t e_t^2.$$

If we expand on both sides by $\mathbb{I}_{A_I} w_t * \mathbb{I}_{A_I} w_t e_t^2$, we get:

$$\begin{aligned}
\mathbb{I}_{A_J} w_t * \mathbb{I}_{A_I} w_t e_t^2 + \mathbb{I}_{A_I} w_t * \mathbb{I}_{A_I} w_t e_t^2 &\leq \mathbb{I}_{A_I} w_t * \mathbb{I}_{A_J} w_t e_t^2 + \mathbb{I}_{A_I} w_t * \mathbb{I}_{A_I} w_t e_t^2, \\
(\mathbb{I}_{A_J} w_t + \mathbb{I}_{A_I} w_t) * \mathbb{I}_{A_I} w_t e_t^2 &\leq \mathbb{I}_{A_I} w_t * \mathbb{I}_{D_T} w_t e_t^2, \\
\frac{\mathbb{I}_{A_I} w_t e_t^2}{\mathbb{I}_{A_I} w_t} &\leq \frac{\mathbb{I}_{D_T} w_t e_t^2}{\mathbb{I}_{D_T} w_t}, \\
\mathrm{WMSE}_{A_i} &\leq \mathrm{WMSE}_{D_T}.
\end{aligned}$$

Hence, the smaller WMSE of the two partitions is also smaller than the WMSE of the whole dataset. Inversely, the larger of the WMSEs of the two partitions is larger than the WMSE of the whole dataset — which can be shown analogously to the provided derivation. □

Proposition S2.4. *Given two partitions, $A_i \subset D_T$ and $A_j \subset D_T$, with $\mathrm{WMSE}_{A_i} \leq \mathrm{WMSE}_{A_j}$, the WMSE_{D_T} is bound by $\mathrm{WMSE}_{A_i} \leq \mathrm{WMSE}_{D_T} \leq \mathrm{WMSE}_{A_j}$.*

S2.4 Low Effort Nash-Sutcliffe Efficiency

For a partition $A_x \in D_T$ the LENSE is defined as:

$$\mathrm{LENSE}_{A_X} = 1 - \frac{\frac{1}{\mathbb{I}_{A_X}} \mathbb{I}_{A_X} e_t^2}{\frac{1}{\mathbb{I}_{X_\mathrm{R}}} \mathbb{I}_{X_\mathrm{R}} e_t^2}, \tag{S8}$$

where X_R defines the data of the reference partition.

Proposition S2.5. *Given two partitions, $A_i \subset D_T$ and $A_j \subset D_T$, with $\mathrm{LENSE}_{A_i} \leq \mathrm{LENSE}_{A_j}$, the LENSE_{D_T} is bound by $\mathrm{LENSE}_{A_i} \leq \mathrm{LENSE}_{D_T} \leq \mathrm{LENSE}_{A_j}$.*

Proof. We rewriting Eq. (S8) in terms of MSE as

$$\mathrm{LENSE}_{A_X} = 1 - \frac{\mathrm{MSE}_{A_X}}{\frac{1}{\mathbb{I}_{X_\mathrm{R}}} \mathbb{I}_{X_\mathrm{R}} e_t^2}. \tag{S9}$$

By inserting Eq. (S9) into the inequality from proposition S2.5 and rearranging we get:

$$\frac{\mathrm{MSE}_{A_i}}{\frac{1}{\mathbb{I}_{X_\mathrm{R}}} \mathbb{I}_{X_\mathrm{R}} e_t^2} \geq \frac{\mathrm{MSE}_{D_T}}{\frac{1}{\mathbb{I}_{X_\mathrm{R}}} \mathbb{I}_{X_\mathrm{R}} e_t^2} \geq \frac{\mathrm{MSE}_{A_j}}{\frac{1}{\mathbb{I}_{X_\mathrm{R}}} \mathbb{I}_{X_\mathrm{R}} e_t^2}. \tag{S10}$$

Here, the denominator is a constant (we can also relate it to the WMSE by observering that $w_t = w = \frac{1}{\frac{1}{\mathbb{I}_{X_\mathrm{R}}} \mathbb{I}_{X_\mathrm{R}} e_t^2}$). Thus, we can reduce the equation to obtain:

$$\mathrm{MSE}_{A_i} \geq \mathrm{MSE}_{D_T} \geq \mathrm{MSE}_{A_j}, \tag{S11}$$

which we know to be true from proof S2.2. □

S2.5 Situation-equitable Nash–Sutcliffe Efficiency

The SENSE for a partition $A_x \in D_T$ is defined as:

$$\mathrm{SENSE}_{A_X} = 1 - \frac{\mathbb{I}_{A_X} \frac{1}{(o_t - \hat{\mu}_t)^2} (o_t - s_t)^2}{\mathbb{I}_{A_X} \frac{1}{(o_t - \hat{\mu}_t)^2}}, \tag{S12}$$

where $\hat{\mu}_t$ is an estimation for the conditional expectation of the observations of at timestep t. Appendix S1 discusses a potential approach for obtaining such an estimation.

Proposition S2.6. *Given two partitions, $A_i \subset D_T$ and $A_j \subset D_T$, with $\mathrm{SENSE}_{A_i} \leq \mathrm{SENSE}_{A_j}$, the SENSE_{D_T} is bound by $\mathrm{SENSE}_{A_i} \leq \mathrm{SENSE}_{D_T} \leq \mathrm{SENSE}_{A_j}$.*

Proof. It is easy to see that the right hand side of Eq. (S12) is a special case of the WMSE with $w = \frac{1}{}$. Thus, the proof is analogous to proof S2.3. □

S2.6 Nash–Sutcliffe Efficiency

The NSE for a partition $A_X \in D_T$ is defined as:

$$\mathrm{NSE}_{A_X} = \frac{\mathbb{I}_{A_X} e_t^2}{\bar{o}_X}. \tag{S13}$$

Proposition S2.7. *Given two partitions, $A_i \subset D_T$ and $A_j \subset D_T$, with $\text{NSE}_{A_j} \geq \text{NSE}_{A_i}$; NSE_{D_T} is bound by below by the smaller* NSE *value of the two partitions — i.e., NSE_{A_i}.*

Proof. Following the convention from S2.3 and Eq. (S13), we get:

$$1 - \frac{\mathbb{I}_{A_J} e_t^2}{\mathbb{I}_{A_J}(o_t - \bar{o}_J)^2} \geq 1 - \frac{\mathbb{I}_{A_I} e_t^2}{\mathbb{I}_{A_I}(o_t - \bar{o}_J)^2}.$$

By subtracting 1, multiplying by -1, and rearranging $\mathbb{I}_{A_J} e_t^2$ and $\mathbb{I}_{A_I} e_t^2$ we get:

$$\frac{\mathbb{I}_{A_J} e_t^2}{\mathbb{I}_{A_I} e_T^2} \leq \frac{\mathbb{I}_{A_J}(o_t - \bar{o}_J)^2}{\mathbb{I}_{A_I}(o_t - \bar{o}_I)^2},$$

adding 1 to both sides and substituting e_t gives:

$$\frac{\mathbb{I}_{A_J} e_t^2}{\mathbb{I}_{A_I} e_t^2} + 1 \leq \frac{\mathbb{I}_{A_J}(o_t - \bar{o}_J)^2}{\mathbb{I}_{A_I}(o_t - \bar{o}_I)^2} + 1,$$

and

$$\frac{\mathbb{I}_{A_J} e_t^2}{\mathbb{I}_{A_I} e_t^2} + \frac{\mathbb{I}_{A_I} e_t^2}{\mathbb{I}_{A_I} e_t^2} \leq \frac{\mathbb{I}_{A_J}(o_t - \bar{o}_J)^2}{\mathbb{I}_{A_I}(o_t - \bar{o}_I)^2} + \frac{\mathbb{I}_{A_J}(o_t - \bar{o}_J)^2}{\mathbb{I}_{A_J}(o_t - \bar{o}_J)^2}. \quad \text{(S14)}$$

At this stage we note that $\bar{o}(A_i)$ minimizes the squared distance for the samples in A_i, and vice versa $\bar{o}(A_j)$ minimizes for samples in A_j. Any other choice than $\bar{o}(A_i)$ (or $\bar{o}(A_j)$, respectively) in their respective partitions leads to a larger sum. For example:

$$\mathbb{I}_{A_J}[o_t - \bar{o}_J)]^2 + \mathbb{I}_{A_I}[o_t - \bar{o}_I)]^2 \leq \mathbb{I}_{A_J}[o_t - \bar{o}_T]^2 + \mathbb{I}_{A_I}[o_t - \bar{o}_T]^2,$$

which we can use to bind the right-hand side of Eq. (S14). After some rearrangement, we have

$$\frac{\mathbb{I}_{D_T} e^2}{\mathbb{I}_{A_I} e^2} \leq \frac{\mathbb{I}_{D_T}[o_t - \bar{o}_T]^2}{\mathbb{I}_{A_I}[o_t - \bar{o}_I)]^2}.$$

From here we can manipulate the equation to reintroduce the canonical formulation of the NSE and we get:

$$1 - \frac{\mathbb{I}_{D_T} e_t^2}{\mathbb{I}_{D_T}(o_t - \bar{o}_T)^2} \geq 1 - \frac{\mathbb{I}_{A_I} e_t^2}{\mathbb{I}_{A_I}(o_t - \bar{o}_I)^2},$$
$$\text{NSE}_{D_T} \geq \text{NSE}_{A_i}.$$

Thus, the lower NSE of the two partitions is the lower bound of the NSE of the whole dataset. □

This means that that we obtain the following relation for the NSE:

$$\text{NSE}_{A_i} \leq \text{NSE}_{D_T} \leq 1.$$

S2.7 Pearson's correlation coefficient

The sample correlation coefficient over a partition $A_X \subseteq D_T$ is

$$r_{A_X} = \frac{\mathbb{I}_{A_X}(o_t - \bar{o}_X)(s_t - \bar{s}_X)}{\sqrt{\mathbb{I}_{A_X}(o_t - \bar{o}_X)^2 \mathbb{I}_{A_X}(s_t - \bar{s}_X)^2}}. \quad \text{(S15)}$$

In most cases, r_{A_z} is independent from $r(D_T)$ — as is demonstrated by Simpson's Paradox (Sect. 1 of the original manuscript) — and we can only make claims for special cases. For example, if $r(D_T) = 1$ (or -1) then it follows that $r_{A_z} = 1$ (or -1).

S2.8 The special case of standardized data

During the review process, reviewer Hoshin Gupta inspired us to think about what would happen to the NSE if the available data would always be standardized (i.e., both the observations and simulations have zero mean and unit variance for all partitions and the overall data). This section shows that in this special setting the NSE and the Kling-Gupta Efficiency (KGE just measures the Pearson's correlation coefficient r, and the correlation becomes the same as the cosine similarity.

Proposition S2.8. *In a setting where we standardize the observations and model outputs for a given set of observations and simulations, we get* $\text{NSE} = 2 * r - 1$.

Proof. As per Gupta et al. (2009) the NSE can be decomposed into

$$\text{NSE} = 2 * \alpha * r - \alpha^2 - \beta, \quad \text{(S16)}$$

where α is the ratio of the standard deviations, i.e.: $\alpha = \frac{\sigma_s}{\sigma_o}$, r is Pearson's correlation coefficient, and $\beta = \frac{\mu_s - \mu_o}{\sigma_o}$ (here independent of the partition).

Since the means of the observations and simulations are zero, it always holds that $\beta = 0$ and $\alpha = 1$, which simplifies Eq. (S16) to

$$\text{NSE} = 2 * r - 1.$$

In other words, in this special setting the NSE only measures the correlation. □

There exists a similar simplification for the KGE:

Proposition S2.9. *In a setting where we standardize the observations and model outputs, it holds that* $\text{KGE} = r$.

Proof. The KGE is defined as

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\sigma_s/\sigma_o - 1)^2 + (\mu_s/\mu_o - 1)^2}.$$

In the current setting $\frac{\mu_s}{\mu_o}$ is actually undefined because of the division by zero, but we might also interpret it as one because $\mu_s = \mu_o$. Similarly, $\frac{\sigma_s}{\sigma_o} = 1$. Thus, the only part within the square root that remains is $(r-1)^2$, which gives us:

$$\begin{aligned}\mathrm{KGE} &= 1 - \sqrt{(r-1)^2}, \\ &= 1 - |r-1|, \\ &= r.\end{aligned}$$

Thus, we showed that in the special setting where observations and simulations are standardized the KGE measures the correlation only. □

Next, we show that within the standardization setting the correlation becomes the cosine similarity.

Proposition S2.10. *In a setting where we standardize the observations and model outputs for all data and all partitions, the correlation is the same as the cosine similarity.*

Proof. The cosine similarity between two N dimensional vectors a and b is defined as

$$s_c = \cos\theta = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2} * \sqrt{\sum_{i=1}^{N} b_i^2}}, \tag{S17}$$

where θ is the angle between the two vectors, and equivalence to r is given because Eq. (S17) is the same as Eq. (S15) if the means are set to zero. □

In other words, we posit that in this special setting the correlation only measures the difference in rotation by the two centered vectors of the observations and simulations.

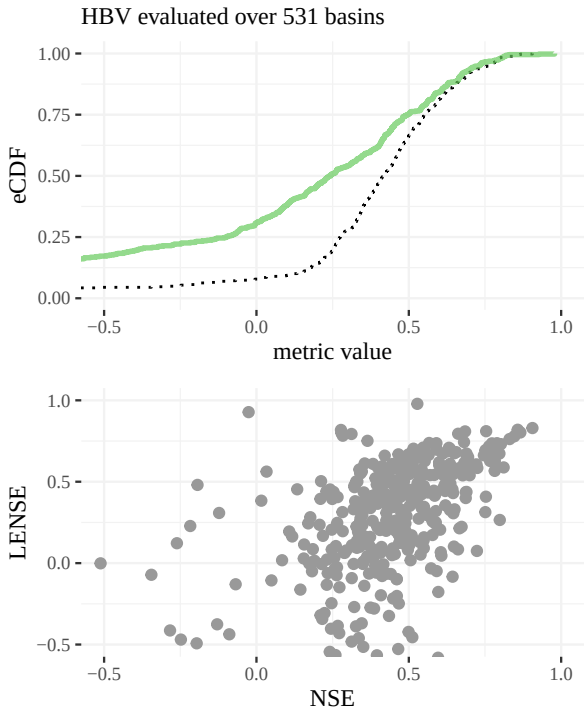

Figure S2. Empirical cumulative distribution functions of the NSE (black, dotted line) and the LENSE (green line) for the 531 CAMELS basins and an ensemble of calibrated HBV models from (see Kratzert et al., 2019).

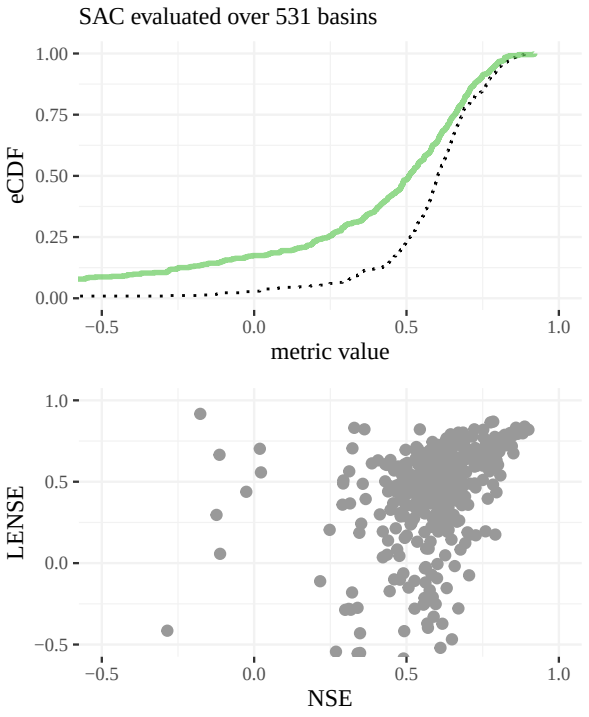

Figure S3. Empirical cumulative distribution functions of the NSE (black, dotted line) and the LENSE (green line) for the 531 CAMELS basins and the SAC-SMA model (see Kratzert et al., 2019).

S3 More experimental results

S3.1 Model evaluation with the NSE and the LENSE for different models

This section shows alterations of the first part of our analysis (Sect. 2.2 of the original manuscript) using different models.

S3.2 Two more experimental results

This appendix shows two additional sweeps of our experiment from Sect.2.1 of the original manuscript. One has $\mathrm{NSE}_{\mathrm{low}} = 0.25$ (Fig. S5) and the other $\mathrm{NSE}_{\mathrm{low}} = 0.75$ (Fig. S6).

S4 Flow–duration curve based metrics and the DAMN.

This appendix provides a short comment on why many of the currently used metrics based on flow–duration curves do not guard against the DAMN. Specifically, we discuss the case of the percent bias of the bottom 30% low flow range (FLV) and percent bias of the top 2% high flow range (FHV) as defined in Yilmaz et al. (2008). Both the FLV and the FHV first divide the data based on the flow-duration curve and then compute the percent bias for the flow–duration values that fall within the predefined partition. This approach has three problems with regard to the DAMN:

1. The a-priori set thresholds (30% and 2%) are too coarse to capture situational differences in model performance. For example, a model that captures rain–driven high-flows well, but melt–driven ones badly might still exhibit a good FHV if the former occur frequently enough to fall over the threshold. This problem is made worse by point 2.

2. The relative bias can be compensated by varying situational performance. For example, a model that overestimates some set of peaks, but equally underestimates another set of peaks can have an FHV that is close to 0 (nearly perfect) — despite the model performing badly for all peaks. This problem is exacerbated by point 3.

3. The flow duration curve breaks temporal locality: The temporal occurrence of the events is not considered, so the behavior in one situation can compensate for the behavior in another. For example, if a model underestimates one high peak A, but overestimates another unrelated event B, then it still can happen that the FHV is be close to zero (i.e., nearly perfect) because the simulations can become assigned to different observations.

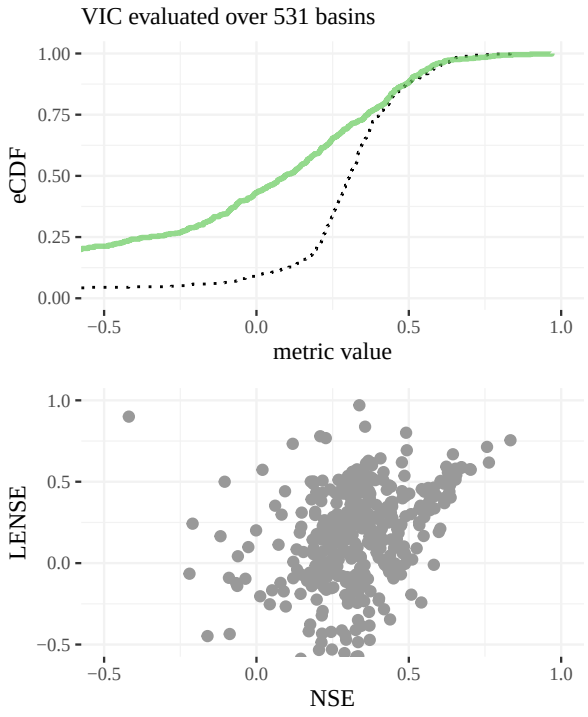

Figure S4. Empirical cumulative distribution functions of the NSE (black, dotted line) and the LENSE (green line) for the 531 CAMELS basins and the VIC-conus model (see Kratzert et al., 2019).

References

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.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, 2019.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water resources research, 44, 2008.

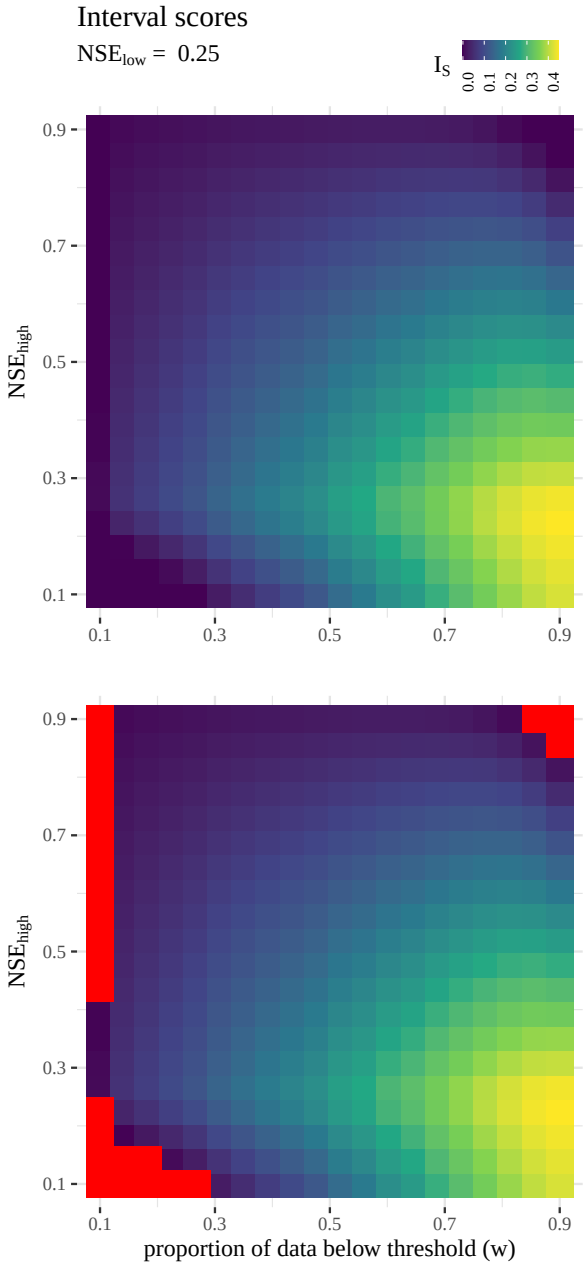

Figure S5. Rerun of our experiment (Sect.2.1 from the original manuscript) with NSE_{low} fixed at 0.25. Each pixel in the plot represents an "interval score", which is zero if the overall model performance NSE_{all} is within the interval spanned by model performance in the low-flow partition, NSE_{low}, and the high-flow partition, NSE_{high}. The red squares in the plot below indicate where where the interval score falls to zero. In the other case, the interval score is negative if NSE_{all} is lower than the interval, and positive if NSE_{all} is higher (see Fig. 4 of the original manuscript).

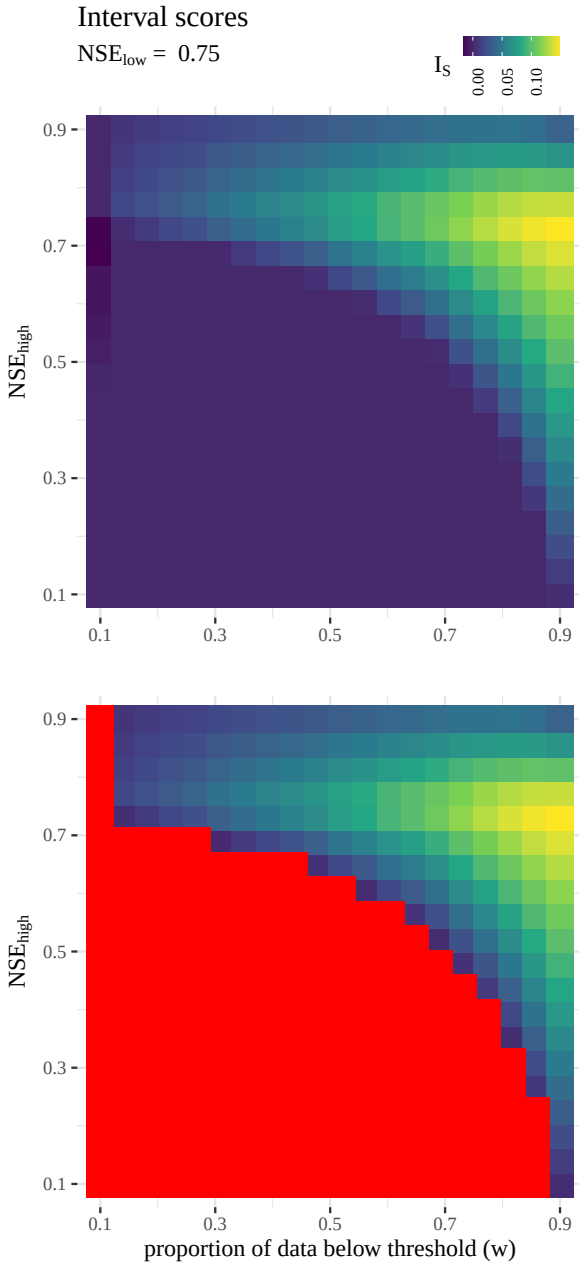

Figure S6. Rerun of our experiment (Sect. 2.1 from the original manuscript) with $\mathrm{NSE}_{\mathrm{low}}$ fixed at 0.75. Each pixel in the plot represents an "interval score", which is zero if the overall model performance $\mathrm{NSE}_{\mathrm{all}}$ is within the interval spanned by model performance in the low-flow partition, $\mathrm{NSE}_{\mathrm{low}}$, and the high-flow partition, $\mathrm{NSE}_{\mathrm{high}}$. The red squares in the plot below indicate where where the interval score falls to zero. In the other case, the interval score is negative if $\mathrm{NSE}_{\mathrm{all}}$ is lower than the interval, and positive if $\mathrm{NSE}_{\mathrm{all}}$ is higher (see Fig. 4 of the original manuscript).