https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

A signal processing-based interpretation of the Nash-Sutcliffe efficiency

Le Duc[12], Yohei Sawada[1]

[1]Institute of Engineering Innovation, University of Tokyo, Tokyo, 113-8656, Japan
[2]Meteorological Research Institute, Tsukuba, 305-0052, Japan

Correspondence to: Le Duc (leduc@sogo.t.u-tokyo.ac.jp)

Abstract. The Nash-Sutcliffe efficiency (NSE) is a widely used score in hydrology but is not common in the other environmental sciences. One of the reasons for its unpopularity is that its scientific meaning is somehow unclear in the literature. This study attempts to establish a solid foundation for NSE from the viewpoint of signal progressing. Thus, a forecast is viewed as a received signal containing a wanted signal (observations) contaminated by an unwanted signal (noise). This view underlines an important role of the error model between forecasts and observations.

By assuming an additive error model, it is easy to point out that NSE is equivalent to an important quantity in signal processing: the signal-to-noise ratio. Moreover, NSE and the Kling-Gupta efficiency (KGE) are shown to be equivalent, at least when there are no biases, in the sense that they measure the relative magnitude of the power of noise to the power of variation of observations. The scientific meaning of NSE explains why it is reasonable to choose NSE=0 as the boundary between skilful and unskilful forecasts in practice, and this has no relation with the benchmark forecast that is equal to the mean of observations. Corresponding to NSE=0, the critical values of KGE is given approximately by 0.5.

In the general cases, when the additive error model is replaced by a mixed adaptive-multiplicative error model, the traditional NSE is shown not to be a well-defined notion. Therefore, an extension of NSE is derived, which only requires to divide the traditional noise-to-signal ratio by the multiplicative factor. This has a practical implication: if the multiplicative factor is not considered, the traditional NSE and KGE underestimate (overestimate) the generalized ones when the multiplicative factor is greater (smaller) than one. In particular, the benchmark forecast turns out to be the worst forecast under the view of the generalized NSE.

1 Introduction

In hydrology, the Nash-Sutcliffe efficiency (NSE) is one of the most widely used similarity measures for calibration, model comparison, and verification (ASCE, 1993; Legates and McCabe, 1999; Moriasi et al., 2007; Pushpalatha et al., 2012; Todini and Biondini, 2017). However, Schaefli and Gupta (2007) pointed out a noticeable fact that NSE is not commonly used even in environmental sciences despite the fact that calibration, model comparison and verification are also employed in such scientific fields. Does this mean that NSE is a special metric that is only relevant for hydrological processes? If this is not the

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

case, what causes this limited use outside of hydrology? One of the reasons can be traced back to the lack of consensual scientific meaning of NSE in literature.

NSE was firstly proposed by Nash and Sutcliffe (1974) by approaching verification from a viewpoint of linear regression (Murphy et el., 1989)

$$NSE = 1 - \frac{\sum(o_i - f_i)^2}{\sum(o_i - \mu_o)^2} = 1 - \frac{\overline{(o-f)^2}}{\overline{(o-\mu_o)^2}}, \qquad (1)$$

where f_i, o_i denote forecasts and observations, respectively, $\overline{(\,)}$ denotes the expectation, and $\mu_o = \bar{o}$ is the mean of observations. The authors noted the analogy between NSE and the coefficient of determination R^2 in linear regression. Since R^2 measures goodness-of-fit in linear regression, NSE should yield a similarity measure for our verification problem. This use of R^2 implies that NSE regresses observations on forecasts

$$o = af + b, \qquad (2)$$

where a, b are the linear regression coefficients, then uses the residual sum $\sum(o_i - f_i)^2$, which they called the residual variance, and the total sum $\sum(o_i - \mu_o)^2$, which they called the initial variance, in the definition of NSE. In general cases, the residual sum should be $\sum(o_i - af_i - b)^2$. This points out that the underlying regression model implicitly assumes the unbiased regression line ($a = 1, b = 0$), which is rarely satisfied in reality.

Identifying NSE to R^2 in linear regression was soon replaced by identifying NSE to skill scores in verification (ASCE, 1993; Moriasi et al., 2007; Schaefli and Gupta, 2007; Ritter and Munoz-Carpena, 2013). Here a skill score measures a relative performance between a score and its benchmark or baseline (Murphy, 1988). This benchmark score is obtained by using a benchmark forecast, which is usually an easily accessible forecast that does not require complicated computation. The most common benchmark forecasts are long-term or climatological forecasts and persistent forecasts. Thus, applying to NSE, the nominator $\sum(o_i - f_i)^2$ is simply the familiar mean squared error score (MSE), while the denominator $\sum(o_i - \mu_o)^2$ is now reinterpreted as the MSE of the benchmark forecast given by the mean of observations $f_i = \mu_o$. Equivalently, NSE can also be viewed as a normalized MSE with the normalizing factor $\sum(o_i - \mu_o)^2$ (Moriasi et al., 2007; Lamontagne et al., 2020).

However, the special choice of μ_o as the benchmark forecast does not somehow conform with the purpose of using skill scores. Here the problem is that, the mean of observations can only be accessed after all observations are realized. This is not available at the time we issue forecasts, and therefore cannot be compared with our forecasts at that time. This subtle problem was noticed by several authors (Legates and McCabe, 1999; Seibert, 2001) and seasonal or climatological means were suggested as benchmarks instead of the mean of observations. However, Legates and McCabe (2012) showed that the appropriate choice of benchmark forecasts depends on hydrological regimes, leading to a more complicated use of NSE in practice. Therefore, they suggested to stick with the original NSE.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

In recent years, starting with the work of Gupta et al. (2009), NSE has been recognized as a compromise of different criteria that measures an overall performance by combining different scores for means, variances and correlations. The decomposition form of NSE in terms of the correlation ρ, the ratio of standard deviations $\alpha = \sigma_f/\sigma_o$, and the ratio of means $\beta = \mu_f/\mu_o$ is given by (Lamontagne et al., 2020)

$$NSE = 2\alpha\rho - \alpha^2 - \frac{(\beta-1)^2}{(\sigma_o/\mu_o)^2}. \qquad (3)$$

Given this unintuitive form of NSE, Gupta et al. (2009) suggested a more intuitive score named the Kling-Gupta efficiency (KGE)

$$KGE = 1 - \sqrt{(\rho-1)^2 + (\alpha-1)^2 + (\beta-1)^2}. \qquad (4)$$

Note that KGE is only one of many potential combinations of ρ, α, β that yield an appropriate verification score. In this multiple-criteria framework, the scientific meaning of a score depends on the skills that we put more weights in consideration. However, unlike KGE, NSE defined by (3) is not a linear combination of the individual scores related to ρ, α, β, therefore, the scientific meaning of NSE is even more obscure in this context. In other words, we can simply explain that NSE measures overall performances, but we cannot separate the contribution from each individual score.

One of weak points of the multiple-criteria viewpoint is that it explains the elegant form (1) by the unintuitive form (3). We suspect that there exists a more profound explanation for the elegant form (1) which also gives us the scientific meaning of NSE. In pursuing this explanation, we will come back to the insight of Nash and Sutcliffe (1974) when they first proposed NSE as a measure. This insight was expressed clearly in Moriasi et al. (2007) when they understood NSE as the relative magnitude of variances of noise and variances of informative signals. This suggests us to approach NSE from the perspective of signal processing. We will show that NSE is indeed a well-known quantity in signal processing.

This paper is organized as follows. In Sect. 2, we revisit the traditional NSE from the viewpoint of signal processing on forecasts and observations. Only with an additive error model imposing on forecasts and observations, the nature and behaviour of NSE in practice can be established. Since the additive error model excludes all forecasts with variances less than variances of observations including the mean forecast, Sect. 3 extends the error model in Sect. 2 by introducing multiplicative biases besides additive biases. Then an extension of NSE in these general cases is derived. Finally, Sect. 4 summarizes the main findings of this study and discusses some implications of using NSE in practice.

2 Specific cases: additive error models

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

2.1 The scientific meaning of NSE

From now on we will consider forecasts and observations from the perspective of signal processing. According to this view, observations form a desired signal that we wish to faithfully reproduce whenever we issue a forecast. Of course, at the time when the forecast is issued, the observations are unknown, and this forecast is the only available signal. This situation is common in signal processing when desired signals are usually unknown and can only be probed through received signals. Therefore, this signal (the forecast) is assumed to be the wanted signal (the observations) contaminated by a certain unwanted signal (noise). This means that we can still get useful information on the observations from the forecast when the noisiness is small. In this section, we assume a simple additive error model for forecasts

$$f = o + b + \varepsilon, \quad (5)$$

where b denotes constant systematic errors, and $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$ denotes random errors with the error variance σ_e^2. The two random variables o and ε are assumed to be uncorrelated.

Using the error model (5), it is easy to calculate two expectations in the formula of NSE

$$MSE = \overline{(f-o)^2} = b^2 + \sigma_e^2, \quad (6)$$

$$\overline{(o-\mu_o)^2} = \sigma_o^2, \quad (7)$$

leading to the following form of NSE

$$NSE = 1 - \frac{b^2+\sigma_e^2}{\sigma_o^2}. \quad (8)$$

The reciprocal of the ratio $(b^2 + \sigma_e^2)/\sigma_o^2$ in (8) recalls the signal-to-noise ratio (SNR) in signal processing

$$SNR = \frac{P_{signal}}{P_{noise}} = \frac{\overline{o^2}}{\overline{(b+\varepsilon)^2}} = \frac{\mu_o^2+\sigma_o^2}{b^2+\sigma_e^2}, \quad (9)$$

where P_{signal}, P_{noise} are the power of the desired signal and noise, respectively. The greater SNR, the better the received signal.

In order to see the relationship between NSE and SNR, we notice that the error model (5) is preserved under the translations $(f, o) \to (f + \Delta, o + \Delta)$ where Δ is an arbitrary real number. This is easy to verify since the same error model is obtained when we add the same value Δ to f, o on both sides of (5). A robust score should reflect this invariance, and therefore is required to be invariant under those translations. If this condition is not satisfied, we will get a different score every time we change the base in calculating water levels for example. It is clear that NSE is translation-invariant while SNR is not. Indeed, we can easily increase SNR by simply increasing μ_o

$$SNR(\Delta) = \frac{(\mu_o+\Delta)^2+\sigma_o^2}{b^2+\sigma_e^2}. \quad (10)$$

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

This is because when a large Δ is added to the desired signal, its magnitude is almost dominated by Δ and the noise magnitude becomes negligible. This suggests that we can use the lower bound of $SNR(\Delta)$, i.e., the SNR in the worst case, as a score to impose the translation-invariant condition

$$SNR_l = \frac{\sigma_o^2}{b^2+\sigma_e^2}. \quad (11)$$

This value is attained when $\Delta = -\mu_o$, which indicates the ratio of the power of variation $o - \mu_o$ to the power of noise. It is worth noting that the translational invariance is violated in the case of KGE since the ratio $\frac{\mu_f+\Delta}{\mu_o+\Delta}$ can vary considerably with Δ.

Since the reciprocal of SNR_l determines NSE in (8), it is more appropriate to define NSE in terms of the noise-to-signal ratio ($NSR = \frac{P_{noise}}{P_{signal}}$):

$$NSE = 1 - \frac{1}{SNR_l} = 1 - NSR_u, \quad (12)$$

where we add the subscript u to NSR to emphasize that this is the upper bound of NSR corresponding to the lower bound of SNR. Thus, under our additive error model, (12) points out that NSE is equivalent to the upper bound of NSR. More exactly, NSE measures the relative magnitude of the power of noise (the unwanted signal) and the power of variation of observations (the wanted signal with its mean removed).

This new interpretation of NSE supports for the choice of the value NSE=0 as the boundary between skilful and unskilful forecasts in practice. In the context of skill scores, NSE of zero indicates the forecasts that yield the same MSE as the mean forecast $f = \mu_o$. A good forecast is therefore required to yield a MSE less than that of the mean forecast. Here in the context of NSR, NSE of zero corresponds to $NSR_u = 1$ or equivalently $\sigma_o^2 = b^2 + \sigma_e^2$. Thus, when NSE tends to go to zero, the background noise tends to have the same power as variation of the desired signal, and consequently corrupts the desired signal. In other words, at NSE=0, we cannot distinguish variation of the observations from noise, and the forecast therefore is useless. Thus, there exist many forecasts yielding the threshold NSE=0, which are not necessarily the mean forecast.

2.2 Random noise-to-signal ratio

Recall that NSE is invariant under the translations along the vector $(1,1)^{\mathrm{T}}$ $(f, o) \to (f + \Delta, o + \Delta)$. However, for general translations $(f, o) \to (f + \Delta_f, o + \Delta_o)$ where the translation vector $(\Delta_f, \Delta_o)^{\mathrm{T}}$ is an arbitrary vector, NSE can take any value

$$NSE = 1 - \frac{(b+\Delta_f-\Delta_o)^2+\sigma_e^2}{\sigma_o^2}. \quad (13)$$

Consequently, we can increase NSE simply by choosing appropriate Δ_f, Δ_o. In practice this approach is known as bias correction with the choice of $\Delta_f \approx -b, \Delta_o = 0$. Since NSE is not invariant under the general translations, misinterpretation

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

on forecast performances can be easily committed. For example, let us consider two forecasts: one with a systematic error and another with a random error

$$f_1 = o + b, \qquad (14a)$$

$$f_2 = o + \varepsilon, \qquad (14b)$$

where we assume $\sigma_o = b = \sigma_e$. Then both f_1, f_2 have $NSE_1 = NSE_2 = 0$, indicating that both forecasts are unskilful. However, it is clear that two forecasts are not equal. A forecaster from his experience knows that the first forecast is better since an almost perfect forecast can be easily obtained from f_1 just by subtracting from f_1 the bias estimated from past observations. In contrast, the performance of f_2 cannot be improved by any translation.

In order to avoid the misjudgement as above, it is desirable to have a score that is invariant under any translation. From (13), it is easy to see that the bias term causes NSE to vary with different displacements of f, o. This motivates us to decompose NSR_u into two components:

$$NSR_u = \frac{b^2+\sigma_e^2}{\sigma_o^2} = \frac{b^2}{\sigma_o^2} + \frac{\sigma_e^2}{\sigma_o^2} = SNSR_u + RNSR_u, \qquad (15)$$

where $SNSR_u$ denotes the systematic NSR_u which changes with the general translations, and $RNSR_u$ denotes the random NSR_u which keeps constant regardless of translations. Thus, $RNSR_u$ is an irreducible component of NSR_u under any translation and acts as a lower bound of NSR_u. Similar to (12) we define a generally invariant version of NSE in terms of $RNSR_u$

$$NSE_u = 1 - \frac{\sigma_e^2}{\sigma_o^2} = 1 - RNSR_u. \qquad (16)$$

Here the subscript u is added to emphasize that this NSE is indeed the upper bound of the original NSE, i.e., the highest NSE can be reached just by translations. NSE_u is identical to NSE when there are no biases in forecasts. For the two forecasts f_1, f_2 in (14a) and (14b), the new score yields $NSE_{u1} = 1, NSE_{u2} = 0$, which reflect our subjective evaluation. As we shall see shortly, $RNSR_u$ will help to ease our analysis on the behaviour of NSE considerably.

We now show an interesting result: $RNSR_u$ can be expressed in terms of a more familiar quantity, the correlation coefficient ρ. This is easy to prove by making use of (5) in the definition of ρ

$$\rho = \frac{\overline{(f-\mu_f)(o-\mu_o)}}{\sigma_f \sigma_o} = \frac{\overline{(o-\mu_o+\varepsilon)(o-\mu_o)}}{\sigma_f \sigma_o} = \frac{\sigma_o^2}{\sigma_f \sigma_o} = \frac{\sigma_o}{\sigma_f}, \qquad (17)$$

and in the definition of σ_f^2

$$\sigma_f^2 = \overline{(f-\mu_f)^2} = \overline{(o-\mu_o+\varepsilon)^2} = \sigma_o^2 + \sigma_e^2. \qquad (18)$$

By plugging (18) into (17), we obtain a one-to-one map between ρ^2 and $RNSR_u$

$$\rho^2 = \frac{\sigma_o^2}{\sigma_f^2} = \frac{\sigma_o^2}{\sigma_o^2+\sigma_e^2} = \frac{1}{1+RNSR_u}, \qquad (19)$$

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

which reveals a profound understanding on ρ, i.e., the correlation reflects noisiness under the error model (5). This is illustrated in Fig. 1 with the joint probability distributions of f, o for different values of ρ.

2.3 Relationships between NSE, KGE, and ρ

The identity (19) enables us to express NSE_u in terms of ρ

$$NSE_u = 1 - RNSR_u = 2 - \frac{1}{\rho^2}, \qquad (20)$$

which is more desirable in mathematical treatment since ρ^2 has support on the finite interval [0,1]. Furthermore, the correlation ρ is a more popular quantity. From (20), it is easy to find the lowest correlation at which a forecast is still considered skilful

$$NSE_u = 2 - \frac{1}{\rho^2} \geq NSE \geq 0 \leftrightarrow \rho \geq \frac{1}{\sqrt{2}} \approx 0.7. \qquad (21)$$

This value can also be deduced from (19) when $RNSR_u = 1$ entails $\rho = 1/\sqrt{2}$. It is worth noting that this critical value of ρ is unknown in the literature.

Similar to NSE_u, we disregard the contribution from the means μ_f, μ_o to KGE in (4) and define its upper bound

$$KGE_u = 1 - \sqrt{(\rho - 1)^2 + \left(\frac{\sigma_f}{\sigma_o} - 1\right)^2} = 1 - \sqrt{(\rho - 1)^2 + \left(\frac{1}{\rho} - 1\right)^2}, \qquad (22)$$

where we have made use of (17) to get the last expression. Recall that although KGE is not invariant under the translations $(f, o) \to (f + \Delta, o + \Delta)$, by excluding the bias term, its upper bound KGE_u now becomes invariant under any translation. It is usually accepted that NSE and KGE do not have a unique relationship, and therefore are not comparable (Konner et al., 2019). However, by focusing on their upper bounds, we can easily compare the two scores on the same plot as depicted in Fig. 2. Several important findings can be drawn from this figure.

Firstly, both scores are monotonic functions of ρ. This is the consequence of the fact that their functional forms are one-to-one maps from ρ to these scores. These bijections ensure that any score ρ, NSE_u or KGE_u can be used as an indirect measure of $RNSR_u$. In this sense, NSE_u, KGE_u, and ρ are only different sides of the same $RNSR_u$, i.e., they are interchangeable in measuring noisiness in forecasts. This points out that KGE has the same scientific meaning as NSE, which indicates the relative magnitude of the power of noise to the power of variation of observations. Although, KGE has been proposed in the multiple-criteria framework, it is interesting to see that the signal processing approach reveals it scientific meaning.

Since we can make any new score simply by assigning any monotonic function of ρ over [0,1] to a score, we illustrate this process for the simplest one that we call correlation efficiency (CE) which is also plotted in Fig. 2

$$CE_u = \rho. \qquad (23)$$

Using (19), we rewrite (23) as

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

$$CE_u = \frac{1}{\sqrt{1+RNSR_u}}. \quad (24)$$

Then replacing $RNSR_u$ by NSR_u, we reintroduce the bias term back into (24) and get the final version

$$CE = \frac{1}{\sqrt{1+NSR_u}} = \frac{\sigma_o}{\sqrt{\sigma_o^2+b^2+\sigma_e^2}}. \quad (25)$$

It is easy to verify that skilful forecasts are indicated by the CE values in the range [0.7,1]. Similarly, we can deduce the translation-invariant form of KGE from (22) by writing ρ in terms of $RNSR_u$, and then replace $RNSR_u$ by NSR_u

$$KGE = 1 - \sqrt{\left(\frac{1}{\sqrt{1+NSR_u}} - 1\right)^2 + \left(\sqrt{1+NSR_u} - 1\right)^2}. \quad (26)$$

Recall that the original KGE (4) is not even invariant under the specific translations $(f, o) \to (f + \Delta, o + \Delta)$. With the new KGE (26) the translational invariance is satisfied. However, replacing $RNSR_u$ by NSR_u is not the only way to enforce the translational invariance, adding a new bias term such as $\frac{b^2}{\sigma_o^2}$ under the square root in (22) also works here.

Secondly, the choice of an appropriate score in practice can be determined by its magnitude and sensitivity. In this sense, Fig. 2 explains why modelers tend to favour KGE in practice. This is because KGE_u is always greater than NSE_u, and at the same time KGE_u is less sensitive to ρ than NSE_u since the derivatives of KGE_u are always smaller than the derivatives of NSE_u. However, in terms of the magnitude and sensitivity of a score, the most appropriate candidate is clearly CE in (25) when CE_u is greater than both KGE_u and NSE_u while its derivatives are constant and smaller than those of KGE_u and NSE_u.

Thirdly, the smaller the correlation, the more sensitive NSE and KGE. This is the consequence of the non-linear dependence of $RNSR_u$ on ρ as expressed in (19). As a result, estimations of KGE and NSE are expected to have high uncertainties when correlations decrease. In contrast, the sensitivity of CE is unaffected with decreasing ρ.

Finally, at the critical value $\rho = 1/\sqrt{2}$ the value of KGE_u is approximately 0.5 (the exact value is $1 - \sqrt{\left(1/\sqrt{2} - 1\right)^2 + \left(\sqrt{2} - 1\right)^2}$), which is the lowest KGE_u at which unbiased forecasts are still considered skilful. It is also the lower bound for the modified KGE (26), which considers all forecasts whether they are biased or not, due to the way it is constructed where $NSR_u \leq 1$ entails $KGE \geq 0.5$. For the traditional KGE (3), the lower bound for good forecasts is not a well-defined concept since this KGE is not just determined by NSR_u. This threshold KGE_*, if exists, has to be equal or greater than 0.5 because otherwise we get a contradiction for unbiased forecasts satisfying $KGE_* < KGE_u < 0.5$. As a result, we come to the necessary condition for a good forecast is that $KGE \geq 0.5$ for any form of KGE.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

Similar to the minimum of $\rho = 1/\sqrt{2}$ for a good forecast, this minimum of KGE is unknown in the literature. In particular, this value is much greater than the corresponding minimum of NSE_u, which is zero. This relatively large gap can lead to misjudgement on forecast performances in practice since similar to NSE, modelers tend to consider KGE=0 as the boundary value between skilful and unskilful forecasts (Anderson et al., 2017; Fowler et al., 2018; Siqueira et al., 2018; Sutanudjaja et al., 2018; Towner et al., 2019). Thus, all forecasts with KGE between [0,0.5] are wrongly classified to be good forecasts while they are indeed unskilful forecasts. It is worth noting that Rogelis et al. (2016) assigned the value KGE=0.5 to be the threshold below which forecasts are considered to be "poor".

The critical value KGE=0.5 is much larger than the KGE value calculated for the case of the mean forecast, which is approximately -0.41 as shown in Knoben et al. (2019). They guessed that -0.41 is the lower bound of KGE for a good forecast. However, we have already seen that both the mean forecast $f = \mu_o$ and the forecasts with $NSR_u = 1$ yield the same NSE=0. How can we explain the different values of -0.41 and 0.5 in the case of KGE? The reason is that the mean forecast does not follow the model error (5). It is clear that the regression line $f = o$ dictated by (5) is very different from the regression line $f = \mu_o$ in the case of the mean forecast. Furthermore, (5) entails that σ_f is always greater than σ_o as shown in (18), which is not the case for $f = \mu_o$. As a result, the error model (5) excludes all forecasts with their variances less than the observation variances, which is expected to commonly occur in practice. This raises the question whether the additive error model holds in reality. If this error model is not followed in reality, can we still use NSE? Another important question is how we introduce the mean forecast into the framework developed so far to examine NSE and KGE? These problems require an extension of the error model (5) and will be further pursued in next section.

3 General cases: mixed additive-multiplicative error models

3.1 Validity of the traditional NSE

In order to extend the additive error model to the general cases, we first notice that the error model (5) indeed gives us the conditional distribution of forecasts on observations. Since all information on forecasts and observations is encapsulated in their joint probability distribution, we can seek the general form of this conditional distribution from their joint distribution in the general cases. For this purpose, we will assume that this joint probability distribution is a bivariate normal distribution

$$\mathrm{p}\begin{pmatrix} f \\ o \end{pmatrix} = \mathcal{N}\left[\begin{pmatrix} \mu_f \\ \mu_o \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_o \\ \rho\sigma_f\sigma_o & \sigma_o^2 \end{pmatrix}\right]. \qquad (27)$$

If the joint distribution is not Gaussian, we need to apply some suitable transformations to f, o such as the root squared transformation $(f, o) \to (\sqrt{f}, \sqrt{o})$, the log transformation $(f, o) \to (\log(f), \log(o))$, the inverse transformation $(f, o) \to (1/f, 1/o)$, … (Pushpalatha et al., 2012). When the joint distribution has the Gaussian form, the conditional distribution also has the Gaussian form (see Chapter 2 in Bishop (2006) for the proof)

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

$$p(f|o) = \mathcal{N}\left[\mu_f + \frac{\rho\sigma_f}{\sigma_o}(o-\mu_o), (1-\rho^2)\sigma_f^2\right]. \quad (28)$$

This implies the following form of the error model

$$f = \frac{\rho\sigma_f}{\sigma_o}o + \left(\mu_f - \frac{\rho\sigma_f}{\sigma_o}\mu_o\right) + \varepsilon = ao + b + \varepsilon, \quad (29)$$

where $a = \frac{\rho\sigma_f}{\sigma_o}$, $b = \mu_f - \frac{\rho\sigma_f}{\sigma_o}\mu_o$, and $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$ with $\sigma_e^2 = (1-\rho^2)\sigma_f^2$. In other words, forecasts in the general cases contain both multiplicative and additive biases besides additive random errors. It is easy to verify that (5) is a special case of (29) when $a = 1$.

It is worth noticing that the nature and behaviour of NSE in Sect. 2 is constructed solely relying on the additive error model without any assumption on the joint probability distribution of (f, o). Therefore, in this section, we again only assume that the error model is described by (29), i.e., a mixed additive-multiplicative error model. The joint distribution is no longer assumed to be a bivariate normal distribution, although (29) is derived from this assumption. This means that the marginal distribution of observations is not restricted to be Gaussian, and can be any probability distribution. However, two important identities obtained with the Gaussian assumption still hold

$$\rho = \frac{\overline{(f-\mu_f)(o-\mu_o)}}{\sigma_f\sigma_o} = \frac{\overline{(ao-a\mu_o+\varepsilon)(o-\mu_o)}}{\sigma_f\sigma_o} = \frac{a\sigma_o^2}{\sigma_f\sigma_o} = \frac{a\sigma_o}{\sigma_f}, \quad (30)$$

$$\sigma_f^2 = \overline{(f-\mu_f)^2} = \overline{(ao-a\mu_o+\varepsilon)^2} = a^2\sigma_o^2 + \sigma_e^2 \rightarrow \sigma_e^2 = (1-\rho^2)\sigma_f^2. \quad (31)$$

Can we now proceed by plugging the error model (29) to the formula (1) of NSE as in Sect. 2? The answer is definitely no because it makes no sense to plug (29) to (1) without first verifying the relevance of the traditional NSE under the error model (29). The most important question is, of course, whether NSE still measures the noise-to-signal ratio. Let us consider three forecasts: two with multiplicative biases and another with a random error

$$f_1 = 2o, \quad (32a)$$

$$f_2 = 0.5o, \quad (32b)$$

$$f_3 = o + \varepsilon, \quad (32c)$$

where we assume $\mu_{f1} = \mu_{f2} = \mu_{f3} = \mu_o = 0$, and $\sigma_o = \sigma_e$. Applying the traditional NSE we obtain $NSE_1 = 0, NSE_2 = 0.75, NSE_3 = 0$, indicating that f_2 is much better than both f_1 and f_3, and the latter have the same skill. But the same in what sense? In the case of f_3 we know that the power of the noise is the same as the power of the observations, rendering $NSE_3 = 0$. Therefore, does the statement that f_1 is as good as f_3 mean that the noise level of f_1 is the same as that of f_3? But f_1 indeed magnifies o by a factor of 2 without any phase error or random error. Furthermore, under the log transformations of f_1 and f_2, these two forecasts show the same magnitude of the additive bias of log2. However, according to NSE, f_2 is much more skilful than f_1. This simple example is enough to show that the scientific meaning of the traditional NSE becomes questionable when we introduce multiplicative biases into the error model.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

We show a further argument for the irrelevance of the traditional NSE under the error model (29) by proving that NSE (1) is not invariant under the translations that preserve the error model (29). In the case of the error model (5) we have shown that this additive error model is preserved under the translations $(f, o) \to (f + \Delta, o + \Delta)$. Geometrically, these translations move the joint distribution along the regression line $f = o + b$. In the general cases (29), the regression line becomes $f = ao + b$. This suggests that the error model (29) is preserved under the translations $(f, o) \to (f + a\Delta, o + \Delta)$, which indeed holds since

$$f + a\Delta = a(o + \Delta) + b + \varepsilon. \quad (33)$$

When $a \neq 1$, these transformations cause NSE (1) to vary with Δ, and therefore the traditional NSE is no longer a robust score under the error model (29).

3.2 An extension of the traditional NSE

In order to seek an appropriate form of NSE in the general cases, we rely on the nature and behaviour of the traditional NSE examined in Sect. 2 by imposing three conditions on the generalized NSE: (1) it measures the noise level in forecasts; (2) it is invariant under the translations $(f, o) \to (f + a\Delta, o + \Delta)$; and (3) its random component, equivalently its upper bound, is invariant under all affine transformations $(f, o) \to (\alpha_f f + \Delta_f, \alpha_o o + \Delta_o)$, where $\alpha_f, \Delta_f, \alpha_o, \Delta_o$ are arbitrary real numbers. Note that we use affine transformations here due to the presence of both multiplicative and additive biases in the error model (29). We proceed by choosing a special transformation, i.e., the bias-corrected transformation $(f, o) \to ((f - b)/a, o)$. This results in an additive error model without biases

$$f_{bc} = \frac{f-b}{a} = o + \frac{\varepsilon}{a}, \quad (34)$$

which suggests that we can define a new NSE in terms of the following upper bound of RNSR

$$RNSR_u = \frac{\sigma_e^2}{a^2\sigma_o^2}. \quad (35)$$

We now prove that (35) is indeed invariant under the transformations $(f, o) \to (\tilde{f} = \alpha_f f + \Delta_f, \tilde{o} = \alpha_o o + \Delta_o)$. In terms of $(\tilde{f}, \tilde{o})$, the error model (29) becomes

$$\tilde{f} = \alpha_f f + \Delta_f = \alpha_f \left(a\frac{\tilde{o}-\Delta_o}{\alpha_o} + b + \varepsilon\right) + \Delta_f = \frac{\alpha_f a}{\alpha_o}\tilde{o} + \alpha_f\left(b - \frac{a\Delta_o}{\alpha_o}\right) + \Delta_f + \alpha_f\varepsilon. \quad (36)$$

Denoting $\tilde{a} = \alpha_f a/\alpha_o, \tilde{\varepsilon} = \alpha_f\varepsilon$, we recalculate (35) for the updated error model (36) with the updated parameters $\tilde{\sigma}_e^2 = \alpha_f^2\sigma_e^2, \tilde{a}^2 = \alpha_f^2 a^2/\alpha_o^2, \tilde{\sigma}_o^2 = \alpha_o^2\sigma_o^2$

$$RNSR_u = \frac{\tilde{\sigma}_e^2}{\tilde{a}^2\tilde{\sigma}_o^2} = \frac{\sigma_e^2}{a^2\sigma_o^2}. \quad (37)$$

Thus, (35) is invariant under any affine transformation, which enables us to define the upper bound of the generalized NSE similar to (16)

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

$$NSE_u = 1 - RNSR_u = 1 - \frac{\sigma_e^2}{a^2\sigma_o^2}. \quad (38)$$

This upper bound entails the desired form of the generalized NSE

$$NSE = 1 - \frac{b^2+\sigma_e^2}{a^2\sigma_o^2} = 1 - \overline{\left(\frac{o}{\sigma_o} - \frac{1}{\rho}\frac{f}{\sigma_f}\right)^2}, \quad (39)$$

where the last expression shows its practical form in comparison with the traditional form (1). We only need to check the invariant property of (39) under the translations $(f, o) \to (f + a\Delta, o + \Delta)$. Since these translations do not alter the bias term b, and are a subset of the affine transformations $(f, o) \to (\alpha_f f + \Delta_f, \alpha_o o + \Delta_o)$, they preserve (39).

In the introduction we have noticed that the decomposition form (3) of NSE is relatively unintuitive even though it is derived from the elegant form (1). From Sect. 3.1 we know that (1) is indeed only relevant under the additive error model (5). It becomes irrelevant when multiplicative biases are introduced into (5). Therefore, if we continue to use the traditional NSE in the general cases, an unintuitive form of NSE will be expected as verified by (3). The appropriate NSE in such cases is the generalized NSE (39).

What is the scientific meaning of the generalized NSE (39)? Clearly, it measures the relative magnitude of the power of noise to the power of variation of observations when the multiplicative factor is removed. Thus, similar to the traditional NSE, the NSE value of zero still marks the boundary between skilful and unskilful forecasts. It also attains the maximum equal to one when forecasts do not have additive biases and random errors. However, there exists a subtle difference in the general cases: the perfect score $NSE = 1$ includes not only the perfect forecast $f = o$, but also all forecasts with only multiplicative biases $f = ao$. This means that this generalized score does not measure the impact of multiplicative biases. In evaluating forecast performances, therefore, we should consider both NSE (39) and the multiplicative factor a although NSE should have a higher priority.

3.3 Behaviour of the generalized NSE

We now prove a surprising result: the upper bound of NSE in the general cases is the same as in the cases of the additive error model, which is given by (20). By making use of the two identities (30), (31) on (38) we have

$$NSE_u = 1 - \frac{\sigma_e^2}{a^2\sigma_o^2} = 1 - \frac{(1-\rho^2)\sigma_f^2}{\rho^2\sigma_f^2} = 2 - \frac{1}{\rho^2}. \quad (40)$$

Thus, in the general cases, correlations still reflect noisiness in forecasts. This is illustrated again in Fig. 3 for the joint probability distributions of f, o with the same $\rho = 0.9$ and different multiplicative factors a. From Fig. 3, it is seemingly counter-intuitive to realize that the noise levels are the same among all forecasts given the same correlations 0.9. Clearly, all the points (f, o) tend to spread wider when increasing a, which implies the noisiness increases. However, this

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

misinterpretation results from our implicit assumption on the additive error model (5) for all the forecasts, i.e., $a = 1$ for all the cases.

A further simple argument will show why this is the case that the noise levels are the same in Fig. 3. Let us consider a forecast $f = o + \varepsilon$. Thus, the simplest way to reduce the magnitude of the random error ε is to multiply f with a very small multiplicative factor a. By doing this, we have a new forecast $\tilde{f} = af$ with a new random error $\tilde{\varepsilon} = a\varepsilon$. Does this mean that $\tilde{f}$ is less noisy than f? Of course, this is not true at all since the noisiness is measured by the relative magnitude between the power of noise and the power of variation of observations, but not by the absolute magnitude of noise. When we multiply f with a, at the same time we multiply o with a, and as a result the relative magnitude is unaltered. This points out further that noisiness of all forecasts $f = ao + a\varepsilon$ for any value of a should be considered to be equivalent. The generalized NSE (39) just reflects this fact.

Since the upper bound of the generalized NSE is invariant when we introduce multiplicative biases into the additive error model (5), all conclusions in Sect. 2.3 still hold. Thus, it is legitimate to use the upper bounds of KGE and CE expressed by (22) and (23), respectively, in the general cases. This implies that the critical values $KGE \approx 0.5$ and $\rho \approx 0.7$ remain to indicate the thresholds below which all forecasts are considered to be unskilful. The generalized CE and KGE can be derived using the same procedure to obtain (25), (26) with the generalized $NSR_u = \frac{b^2+\sigma_e^2}{a^2\sigma_o^2}$ in place of the traditional $NSR_u = \frac{b^2+\sigma_e^2}{\sigma_o^2}$. We derive the generalized CE for illustration

$$CE = \frac{1}{\sqrt{1+NSR_u}} = \frac{a\sigma_o}{\sqrt{a^2\sigma_o^2+b^2+\sigma_e^2}}. \qquad (41)$$

It is worth noting that the form (22) of KGE_u when rewritten using the error model (29) will replace the variance term $\left(\frac{\sigma_f}{\sigma_o} - 1\right)^2$ in the traditional KGE by $\left(\frac{\sigma_f}{a\sigma_o} - 1\right)^2$

$$KGE_u = 1 - \sqrt{(\rho-1)^2 + \left(\frac{1}{\rho} - 1\right)^2} = 1 - \sqrt{(\rho-1)^2 + \left(\frac{\sigma_f}{a\sigma_o} - 1\right)^2}. \qquad (42)$$

Combined with the generalized NSE_u (38), we see that in practice if a is not taken into account, i.e., the traditional NSE and KGE are still used, we underestimate (overestimate) NSE and KGE when $a > 1$ ($a < 1$).

With the generalized NSE, it is now possible to deal with the case of the mean forecast $f = \mu_o$. We exclude the trivial case $o = f = \mu_o$ and always assume $\sigma_o \neq 0$. The mean forecast is equivalent to the following model error

$$f = 0 * o + \mu_o + 0, \qquad (43)$$

which implies $a = 0, b = \mu_o, \sigma_e = 0$ in (29). This specific error model highlights a problem that we have omitted when defining the generalized NSE: $RNSR_u$ (35), and therefore NSE (39), can only be defined for the cases $a \neq 0$. When $a = 0$,

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

forecasts and observations are two uncorrelated signals ($\rho = 0$), and it makes no sense to state that the received signal (forecasts) is the true signal (observations) contaminated by noise.

In order to assign an appropriate value of NSE for the cases $\rho = 0$, we rely on the continuity of NSE_u with respect to ρ as shown in (40). Let ρ approach zero in (40), we get the limit $NSE_u = -\infty$. Since NSE_u is the upper bound of NSE, this entails $NSE = -\infty$. The same argument also yields $KGE = -\infty$ under the limit $\rho \to 0$. In other words, all forecasts uncorrelated to observations, which include the mean forecast, should be classified to the worst forecast with $NSE = -\infty$. It can be justified by noticing that information on variation of observations is totally unknown if only available is an uncorrelated forecast. The generalized NSE therefore provides a new interpretation of the mean forecast. Rather than a benchmark marking the boundary between skilful and unskilful forecasts, the mean forecast is indeed the worst forecast which can be beat by any forecast correlated to observations.

4 Conclusion

The Nash-Sutcliffe efficiency is a widely used score in hydrology but is not common in the other environmental sciences. One of the reasons for its unpopularity is that its scientific meaning is somehow unclear in the literature. There exist many attempts to establish a solid foundation for NSE from several viewpoints: linear regression, skill scores, multiple-criteria scores. This study contributes to these studies by approaching NSE from the viewpoint of signal progressing. Thus, a forecast is viewed as a received signal containing a wanted signal (observations) contaminated by an unwanted signal (noise). This view underlines an important role of the error model between forecasts and observations, which is usually implicit in our assumption.

By assuming an additive error model, it is easy to point out that NSE is equivalent to an important quantity in signal processing: the signal-to-noise ratio. More precisely, NSE measures the relative magnitude of the power of noise to the power of variation of observations. Therefore, NSE is a universal metrics that should be applicable in any scientific fields. Its scientific meaning explains why it is reasonable to choose NSE=0 as the boundary between skilful and unskilful forecasts in practice. This is because when NSE goes below zero, the power of noise starts dominating the power of variation of observations, meaning that noise distorts the desired signal and make it difficult to extract the useful information. This choice has no relation with the interpretation that NSE=0 corresponds to the benchmark forecast that is equal to the mean of observations, and all good forecasts need be better than this simple benchmark.

Since NSE can be easily increased simply by adding some appropriate values into forecasts and observations, we seek its upper bound NSE_u under all such translations. NSE_u is pointed out to correspond to the random component of NSR, and is a useful concept in analysing the behaviour of not only NSE but also KGE. It turns out that NSE_u and KGE_u are different

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

measures of the same RNSR, which can be mapped one-to-one. More surprising, it is found that NSE_u and KGE_u in their turn can be expressed in terms of a more familiar quantity: the correlation. This implies that NSE_u and KGE_u do not introduce any new score, and can equivalently be replaced by ρ. In this sense, any new score can be constructed from ρ with any monotonic function of ρ. This leads to an important finding: corresponding to NSE=0, $\rho \approx 0.7$ and $KGE \approx 0.5$ (not KGE=0) mark the boundary between skilful and unskilful forecasts, which has a practical implication on the use of KGE since modelers usually identify KGE=0 as this boundary similar to NSE=0. Thus, forecasts with KGE between 0-0.5 can be wrongly classified as good forecasts in practice.

Since the additive error model disregards all forecasts with their variances less than observation variances including the mean forecast, we need to work with a more general error model to deal with such cases. By assuming a bivariate normal distribution between forecasts and observations, the general error model is found to be the mixed additive-multiplicative error model. Then under the general cases the traditional NSE is shown not to be a well-defined notion. Therefore, an extension of NSE need be derived. By requiring that the generalized NSE is invariant under affine transformations of forecasts and observations induced by the general error model, it is found to be the traditional one adjusted by the multiplicative factor. Again, this has a practical implication on the use of NSE and KGE: if the multiplicative factor is not taken into account and the traditional ones are used instead, both the scores are underestimated (overestimated) when the multiplicative factor is greater (smaller) than one. The critical values NSE=0, $KGE \approx 0.5$, $\rho \approx 0.7$ still hold with the generalized NSE and KGE.

Finally, we summarize here some profound explanations that the signal processing approach to NSE proposes

- Despite their different forms, NSE and KGE are equivalent, at least when there are no biases, in the sense that they measure the noise-to-signal ratio between the power of noise and the power of variation of observations.
- The critical value NSE=0 that marks the boundary between skilful and unskilful forecasts has no relation with the mean forecast $f = \mu_o$. Its choice is dictated by the fact that at this value the power of noise starts dominating the power of variation of observations.
- Corresponding to NSE=0, the critical values of KGE and the correlation coefficient are given approximately by 0.5 and 0.7, respectively.
- The traditional form of NSE only reflects the noise-to-signal ratio under the additive error model. It no longer reflects this when multiplicative biases are introduced, and as a result has an unintuitive form in the general cases.
- It is necessary to adjust the traditional NSE in the general cases to take into account the effect of multiplicative biases on the noise-to-signal ratio. If this effect is not considered and the traditional one continues to be used, NSE is underestimated (overestimated) when the multiplicative factor is greater (smaller) than one.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

- All forecasts uncorrelated to observations are considered to be the worst forecast when measured by NSE or KGE because no information on variation of observations can be retrieved in these cases. The mean forecast $f = \mu_o$ belongs to this class of forecasts. Therefore, in the view of NSE the mean forecast should not be used as a benchmark forecast.

Author contribution. LD raised the idea and prepared the manuscript. The idea has further been developed in discussion. YS corrected the treatment of NSE and KGE in hydrology and revised the manuscript.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgments. This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) as the Program for Promoting Researches on the Supercomputer Fugaku, Large Ensemble Atmospheric and Environmental Prediction for Disaster Prevention and Mitigation (hp200128, hp210166, hp220167), Foundation of River & basin Integrated Communications (FRICS), and JST Moonshot R&D project (grant no. JPMJMS2281).

References

Andersson, J. C. M., Arheimer, B., Traoré, F., Gustafsson, D., and Ali, A.: Process refinements improve a hydrological model concept applied to the Niger River basin. *Hydrol. Process.*, **31**, 4540–4554, doi:10.1002/hyp.11376, 2017.

ASCE: Criteria for evaluation of watershed models. *J. Irrigation Drainage Eng.,* **119**, 429-442, 1993.

Bishop, C. M.: *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., Woods, R., and Zhang, L.: Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement. *Water Resour. Res.*, **54**, 9812–9832, doi:10.1029/2018WR023989, 2018.

Gupta, H. V., and Kling, H.: On typical range, sensitivity and normalization of mean squared error and Nash-Sutcliffe efficiency type metrics. *Water Resources Res.*, **47**, W10601, doi:10.1029/2011WR010962, 2011.

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

Lamontagne, J. R., Barber, C. A., and Vogel, R. M.: Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Res.*, **56**, e2020WR027101, doi:10.1029/2020WR027101, 2020

Legates, D. R., and McCabe., G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Res.,* **35**, 233-241, doi:10.1029/1998WR900018, 1999.

Legates, D.R., McCabe, G.J.: Short communication a refined index of model performance. A rejoinder. *Int. J. Climatol.* doi:10.1002/joc.3487, 2012.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, **50**, 885–900, doi:10.13031/2013.23153, 2007.

Murphy, A.: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2, 1988.

Murphy, A.H., Brown, B.G. and Chen, Y-S.: Diagnostic verification of temperature forecasts. *Weather Forecast.*, **4**, 485–501, 1989.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models. Part 1: A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Pushpalatha, R., Perrin, C., Le Moine, N., and Andreassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, **420**, 171–182, 2012.

Ritter, A., and Munoz-Carpena, R.: Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, **480**(3), 33–45, doi:10.1016/j.jhydrol.2012.12.004, 2013.

Rogelis, M. C., Werner, M., Obregón, N., and Wright, N.: Hydrological model assessment for flood early warning in a tropical high mountain basin. *Hydrol. Earth Syst. Sci. Discuss.*, doi:10.5194/hess-2016-30, 2016.

Schaefli, B., and Gupta, H. V.: Do Nash values have value? *Hydrol. Processes*, **21**, 2075–2080, doi:10.1002/hyp.6825, 2007.

Seibert, J.: On the need for benchmarks in hydrological modelling. *Hydrol. Process.,* **15**(6), 1063–1064, doi:10.1002/hyp.446, 2001.

Siqueira, V. A., Paiva, R. C. D., Fleischmann, A. S., Fan, F. M., Ruhoff, A. L., Pontes, P. R. M., Paris, A., Calmant, S., and Collischonn, W.: Toward continental hydrologic–hydrodynamic modeling in South America. *Hydrol. Earth Syst. Sci.*, **22**, 4815–4842, doi:10.5194/hess-22-4815-2018, 2018.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geosci. Model Dev.*, **11**, 2429–2453, doi:10.5194/gmd-11-2429-2018, 2018.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., and Stephens, E. M.: Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin. *Hydrol. Earth Syst. Sci.*, **23**, 3057–3080, doi:10.5194/hess-23-3057-2019, 2019.

Todini, E., and Biondi, D.: Calibration, parameter estimation, uncertainty, data assimilation, sensitivity analysis, and validation. *Chap 22 in Handbook of applied hydrology* (pp. 22-1–22-19). New York: McGraw Hill, 2017.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

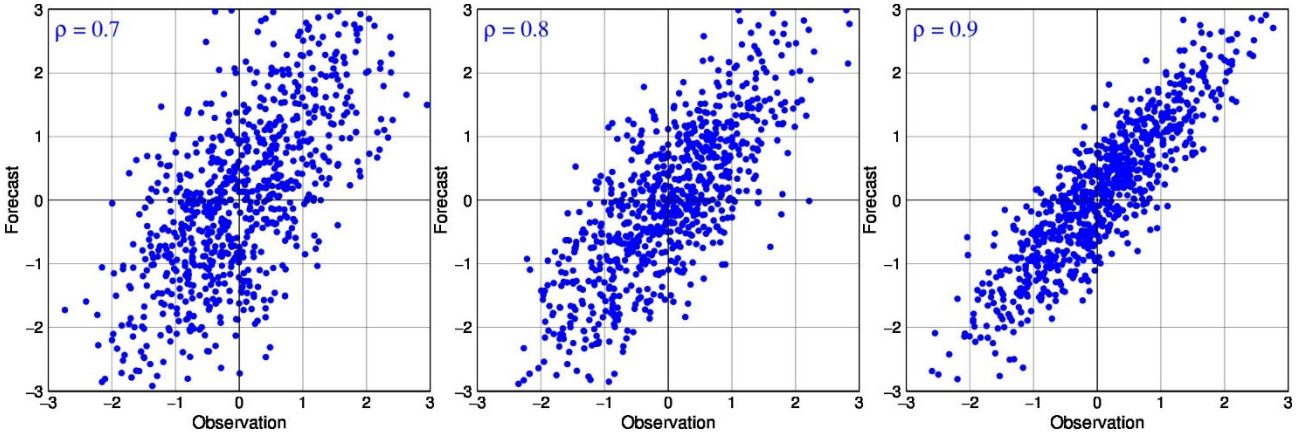

Figure 1: Joint probability distributions of forecasts and observations with different values of ρ under the additive error model.

Here we assume $b = 0$, $o \sim \mathcal{N}(0, 1)$, then the error model yields $\sigma_e = \sqrt{\frac{1}{\rho^2} - 1}\sigma_o$.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

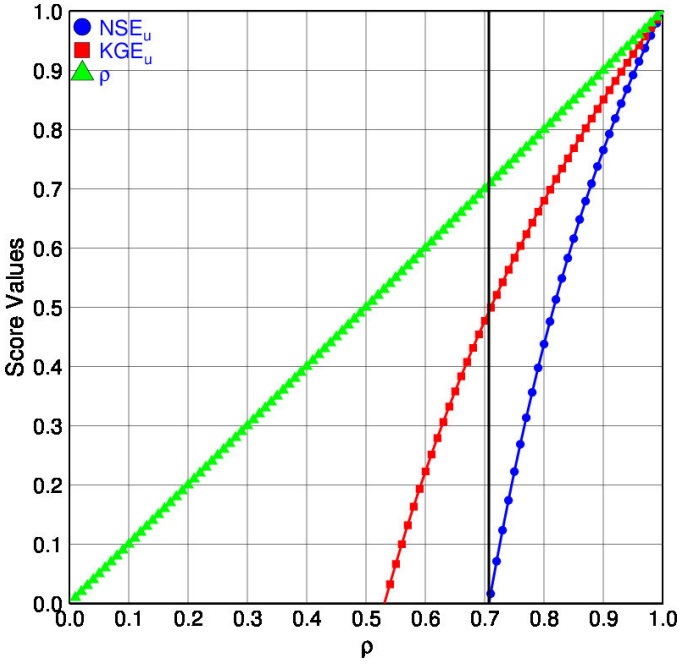

Figure 2: The upper bounds of NSE and KGE as the functions of ρ. The solid line without symbols marks the boundary between unskilful forecasts on the left, and skilful forecasts on the right if forecasts have no biases. If there exist biases in forecasts, this boundary will shift to the right.

https://doi.org/10.5194/egusphere-2022-955
Preprint. Discussion started: 23 September 2022

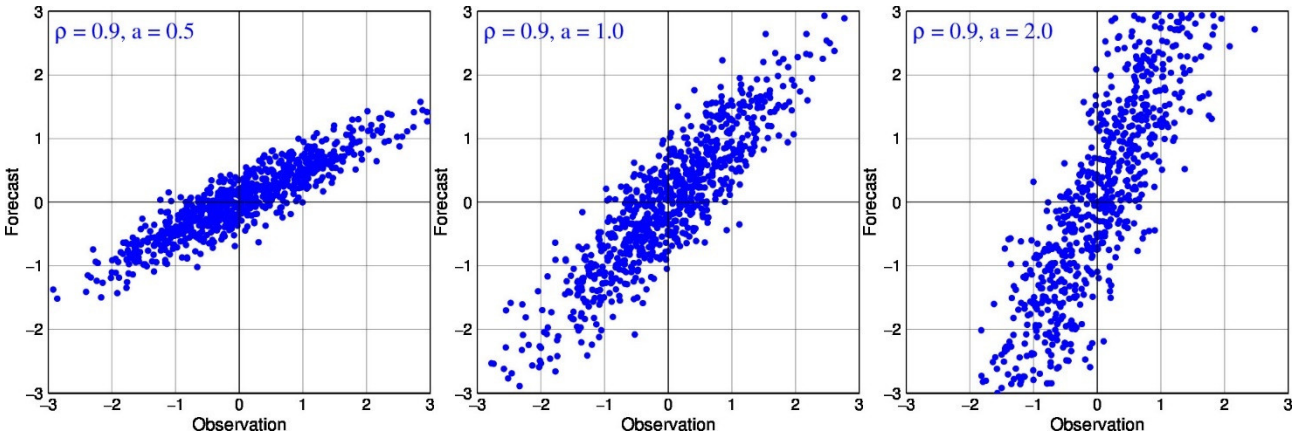

Figure 3: Joint probability distributions of forecasts and observations with the same $\rho = 0.9$ and different values of a under the multiplicative and additive error model. Here we assume $b = 0$, $o \sim \mathcal{N}(0, 1)$, then the error model yields $\sigma_e = \sqrt{\frac{1}{\rho^2} - 1}\, a\sigma_o$. The noise levels as measured by the generalized NSE indicate the same noisiness for all forecasts, even though the noisiness seemingly increases with increasing a.