Check for updates

A hydrological model skill score and revised R-squared

Charles Onyutha IWA (iD)

Department of Civil and Environmental Engineering, Kyambogo University, P.O. Box 1, Kyambogo, Kampala, Uganda
E-mail: conyutha@kyu.ac.ug

(iD) CO, 0000-0002-0652-3828

ABSTRACT

Despite the advances in methods of statistical and mathematical modeling, there is considerable lack of focus on improving how to judge models' quality. Coefficient of determination (R^2) is arguably the most widely applied 'goodness-of-fit' metric in modelling and prediction of environmental systems. However, known issues of R^2 are that it: (i) can be low and high for an accurate and imperfect model, respectively; (ii) yields the same value when we regress observed on modelled series and vice versa; and (iii) does not quantify a model's bias. A new model skill score E and revised R-squared (RRS) are presented to combine correlation, bias measure and capacity to capture variability. Differences between E and RRS lie in the forms of correlation and variability measure used for each metric. Acceptability of E and RRS was demonstrated through comparison of results from a large number of hydrological simulations. By applying E and RRS, the modeller can diagnostically identify and expose systematic issues behind model optimizations based on other 'goodness-of-fits' such as Nash–Sutcliffe efficiency (NSE) and mean squared error. Unlike NSE, which varies from $-\infty$ to 1, E and RRS occur over the range 0–1. MATLAB codes for computing E and RRS are provided.

Key words: distance correlation, hydrological models, model performance evaluation, Nash–Sutcliffe efficiency, revised R-squared (RRS), R-squared

HIGHLIGHTS

- R^2 is arguably the most widely applied 'goodness-of-fit' measure.
- R^2 has known issues e.g. it (i) does not quantify bias, (ii) can be low and high for an accurate and imperfect model, respectively.
- Revised R^2 (RRS) and a metric E are presented to address the issues of R^2.
- E & RRS allow diagnostic exposure of systematic issues behind model optimizations based on other 'goodness-of-fits' such as mean squared error.

INTRODUCTION

Given the growing concern about the impacts of changing climate on hydrology, several studies have been conducted on modelling of hydrological systems. Despite the advances in methods of statistical and mathematical modelling, Alexander *et al.* (2015) asserted that there continues to exist substantial inattentiveness to improving ways to judge the quality of models. Here, the word quality is analogous to how well a model fits through points of measurements or observations and this can be described as 'goodness-of-fit'. In Table 1 of the paper by Blöschl *et al.* (2019), issues 19 and 20 of the 23 unsolved problems in hydrology concern modelling methods. These issues are regarding the adaptation of hydrological models for making extrapolations, and the need for disentangling and reducing model uncertainty in hydrological prediction. It is intuitive that the acceptability of a model's predictability is synonymous with the model's performance or quality. In this way, there are a large number of recent studies which have focused on metrics for assessing model performance and examples of the relevant papers include Bai *et al.* (2021), Clark *et al.* (2021), Stoffel *et al.* (2021), Ye *et al.* (2021), Lamontagne *et al.* (2020), Liu (2020), Barber *et al.* (2019), Jackson *et al.* (2019), Mizukami *et al.* (2019), Rose & McGuire (2019), Pool *et al.* (2018), Lin *et al.* (2017) and Jie *et al.* (2016). Due to efforts to improve how to judge model performance, several 'goodness-of-fit' metrics exist in the literature. However, Jackson *et al.* (2019) suspect that some measures of 'goodness-of-fits' are mostly preferred to others because of familiarity, without focus on the strengths and weakness of the metrics. In fact, a modeller is required to have comprehensive knowledge of the strengths and limitations of a particular 'goodness-of-fit' metric before using it to calibrate a hydrological model (Mizukami *et al.* 2019; Ferreira *et al.* 2020).

Lamontagne *et al.* (2020) noted that, nowadays, the most widely used 'goodness-of-fits' for assessing model performance in hydrology include the Nash–Sutcliffe efficiency (NSE) (Nash & Sutcliffe 1970) and the Kling–Gupta efficiency (Gupta *et al.* 2009). However, the coefficient of determination (hereinafter denoted as R^2 or interchangeably used with R-squared) was known perhaps to be solely the most extensively applied measure of 'goodness-of-fit' (Kvålseth 1985) and this still holds true until the time of writing this paper, especially when we consider the application of R^2 in hydrological, ecological, agricultural and climatic categories of the environmental systems.

Specifically in hydrology, R^2 is regarded as a standard metric for the evaluation of the 'goodness-of-fit' between model simulations and observations (Barber *et al.* 2019). A high value of R^2 tends to be associated with an effective model (Quinino *et al.* 2013). A model performance for which R^2 is zero is intuitively poor. On an important note, there are strong statements in the literature about the use of R^2 for model performance evaluation. For instance, it was remarked that 'the most important aspect of R^2 is that it has no importance in the classical model of regression' (Goldberg 1991). Furthermore, Cameron (1993) warned that R^2 is not a statistical test, and there seems to be no intuitive justification for its use as a descriptive statistic. This suggests that the value of R^2 should not even be reported (Quinino *et al.* 2013) for modelling studies. Correlation from which we obtain R^2 is unsuited for analyses of agreement between observations and model simulations (Schober *et al.* 2018). Furthermore, there are a number of common misunderstandings surrounding the use of R^2 as a measure of model fit, thereby making the application and interpretation of R^2 significantly confusing (Alexander *et al.* 2015). The use of R^2 may not be 'morally ethical' in some contexts (Rose & McGuire 2019). There are quite a number of reasons which render predictive power of R^2 to be inadequate. First of all, R^2 can be low for an accurate model and, on the other hand, an inaccurate model can yield high R^2. Second, we obtain the same value of R^2 by regressing observed (X) and modelled (Y) series and vice versa. This invalidates the use of R^2 as the coefficient of determination to indicate the amount of the total variance in observations explained by the model. Third, R^2 does not quantify bias in a model. Although R^2 is commonly used, its values do not change with respect to magnitude scale change (Jackson *et al.* 2019). It is also known that R^2 is sensitive to outliers in the sample (Barber *et al.* 2019). The conventional R^2 provides invalid results in the presence of measurement errors in the data (Cheng *et al.* 2014). Despite these issues regarding the use of R^2, there are many recent studies which can be found to have applied R^2 for the evaluation of model performance. The metric R^2 (along with gradient and intercept of the corresponding regression line) was recommended by Moriasi *et al.* (2015) as one of the 'goodness-of-fit' metrics for evaluating model performance. In fact, some scientists even advocate for the preference of R^2 to other 'goodness-of-fit' measures of model performance (see, e.g., Chicco *et al.* 2021). Nevertheless, R^2 should not be used for model performance evaluation despite its wide use and recommendation by some scientists (Jenkins & Quintana-Ascencio 2020). The need to introduce a 'goodness-of-fit' metric which addresses a number of (if not all the) issues of R^2 comprises the rationale of this study.

It is worth noting that R^2 has several variants used to measure 'goodness-of-fits'. At least nine R^2 versions can be found elaborated by Kvålseth (1985). Four most common versions of R^2 include: (i) Nash–Sutcliffe efficiency (NSE) (Nash & Sutcliffe 1970), (ii) sum of squared deviations of modelled data points from their mean divided by the sum of squared deviations of observed data points from their mean, (iii) fraction of the variance in observed data explained by the model, and (iv) squared Pearson product moment correlation coefficient. The R^2 formula given by Equation (3) in a paper by Nash & Sutcliffe (1970) can be compared with the R^2 or Equation (1) of Kvålseth (1985). In summary, R^2 versions (i)–(iv) above can be found given by Kvålseth (1985) as R_1^2, R_2^2, R_4^2 and R_6^2 in Equations (1), (2), (4) and (6), respectively. For an ideal model (where the bias is zero), R^2 versions (i) and (iii) are identical to each other. Furthermore, R_4^2 is very close to NSE for any reasonable model (with small bias). However, R_4^2 will always be larger than NSE if the bias is greater than zero. While there is the growing use of R_6^2 or the most well-known version of R^2 for assessing model performance in hydrology despite the past warnings by some scientists such as Goldberg (1991) and Cameron (1993), many researchers when referring to R^2 (or the coefficient of determination) can actually be found using NSE. However, it is important to note that R_6^2 and NSE are different (Bardsley 2013). For instance, when we swap X and Y series, R_6^2 remains unchanged, something which is not the case for NSE. Generally, the best form of R-squared appears to be what modellers (especially hydrologists) refer to as the NSE (Kvålseth 1985). Eventually, NSE has been favoured for assessing 'goodness-of-fits' especially in hydrology. However, R_6^2 or R^2 is so widely used not only in hydrology but also for modelling and prediction of other environmental systems including ecological, agricultural and climatic fields. The formula for NSE is given by (Nash & Sutcliffe 1970):

$$NSE = 1 - \frac{\sum (X - Y)^2}{\sum (X - \bar{X})^2} \tag{1}$$

where $\bar{X}$ is the mean of the Xs and all summations are from $i = 1$ to $i =$ sample size (n) while the subscript i was omitted to simplify the notation, and this style can be found adopted and hereinafter consistently used throughout the paper to avoid any ambiguities.

NSE has a number of issues despite its popularity and wide usage in hydrological modelling. This is why the suitability of NSE has been on the modellers' radar for decades, as this can be seen, for instance, in Garrick *et al.* (1978), Kvålseth (1985), Legates & McCabe (1999, 2013), Krause *et al.* (2005), Gupta *et al.* (2009), Liu (2020) and Clark *et al.* (2021). Eventually, there are several variants of NSE based on its modifications to address issues related to the use of the original version proposed by Nash & Sutcliffe (1970). For instance, a modification of NSE or R^2 in terms of R_9^2 (see Equation (11) from the paper by Kvålseth (1985)), comprised medians M of absolute values of residuals and magnitudes of deviations of the Xs from $\bar{X}$. This R_9^2, which Kvålseth (1985) termed as resistant R^2 statistic (hereinafter denoted as RSS), was to be taken as a measure of the proportion of the total variability of X explained (accounted for) by the fitted model and can be computed using $RSS = 1 - (M(|X - Y|)/M(|X - \bar{X}|))^2$. Noticeably, RSS can be far lower than zero, especially when $M(|X - Y|) \gg M(|X - \bar{X}|)$. Another improvement of NSE was by Legates & McCabe (1999), in terms of the ratio of the sum of absolute (instead of squared) differences between X and Y divided by the sum of absolute (and again not the squared) deviations of the Xs from $\bar{X}$. Furthermore, to overcome the oversensitivity of NSE to peak high values stemming from the influence of squaring the error term, logarithmic NSE is also sometimes used (Krause *et al.* 2005). Despite these improvements for NSE, the original version by Nash & Sutcliffe (1970) continues to be more applicable than its variants. The popularity of the original NSE version could be due to its simplicity compared with the variants. Nevertheless, NSE values occur over a wider range (from $-\infty$ to positive 1) than the expected typical relative error measure with standard values of zero and one characterizing imperfect and perfect model, respectively. Additionally, NSE directly relies on $\bar{X}$ as the comparison baseline and in this way it can lead to overestimation of model skill when modelling highly seasonal river flow, for instance, from runoff in snowmelt-dominated basins (Gupta *et al.* 2009). Moreover, the probability distribution of squared errors between model simulations and observations has heavy tails, thereby making the sampling uncertainty in NSE estimator substantial (Clark *et al.* 2021).

In a relevant development regarding hydrological modelling, performance evaluation has been based on multiple criteria. However, the use of several criteria in a single calibration complicates the application and automation of well-known search algorithms, including generic algorithm and uniform random search as well as many others. Alternatively, the various criteria can be combined into a single 'goodness-of-fit' metric. To do so, it is vital to analyse the similarities and differences between the various criteria before their possible combination. The idea here is that criteria which are not mathematically related could be combined. In this way, we minimize redundancy of related criteria during model calibration.

Therefore, the purpose of this study was to revisit R-squared and also introduce a model skill score in line with the need to have a metric which can address the known issues of R-squared. Here, the rationale for introducing a new model performance metric is to characterize model performance in terms of measures which can enhance the modeller's understanding of the issues behind poor model performance. To allow a modeller to diagnostically identify and expose systematic issues behind unrealistic model outputs, the new metric comprises components that can be used to determine the criteria to which the poor or bad model performance should be attributed. The remaining parts of this paper are organized as follows. The section of materials and methods contains an overview of the previous decompositions of NSE, derivation of the hydrological model skill score and the new formula after revisiting the R-squared. This section also consists of the application of the new model skill score for a case study as well as comparison with other 'goodness-of-fit' measures. The next section comprises results and discussion of the case study. Finally, conclusions are drawn in the final section before the list of references.

MATERIALS AND METHODS

Previous decompositions of NSE

From Equation (1), it can be noted that the numerator and denominator of the second part of the NSE are related to the mean squared error (MSE) and sample variance of X (or s_X^2), respectively, such that $NSE = 1 - (MSE/s_X^2)$. Previously, there have been a few studies (Yates 1982; Murphy 1988; Gupta *et al.* 2009) which comprised decompositions of MSE or NSE. Consider s_Y^2 as the sample variance of Y while $\bar{Y}$ denotes the mean of the Ys. Furthermore, let s_{YX} be the sample covariance of X and Y. MSE can be decomposed using $MSE = (\bar{Y} - \bar{X})^2 + s_Y^2 + s_X^2 - 2s_{YX}$ (Yates 1982). In Murphy (1988), NSE was decomposed into three components, such that $NSE = r^2 - (r^2 - s_Y \times (s_X)^{-1})^2 - ((\bar{Y} - \bar{X}) \times (s_X)^{-1})^2$ where r is the coefficient of Pearson correlation between X and Y while s_X and s_Y represent the sample standard deviation of X and Y, respectively. The first, second and third parts of the decomposed NSE quantify correlation strength, conditional bias and unconditional bias, respectively (Murphy 1988).

An alternative decomposition of NSE was in terms of Kling–Gupta efficiency (KGE) (Gupta *et al.* 2009), such that:

$$KGE = 1 - \sqrt{(r^2 - 1)^2 + \left(\frac{s_Y}{s_X} - 1\right)^2 + \left(\frac{\bar{Y}}{\bar{X}} - 1\right)^2} \qquad (2)$$

KGE is implicitly based on the assumptions of data normality and the absence of outliers (Pool *et al.* 2018). While maximizing KGE, values (the means of simulation) that underestimate $\bar{X}$, especially in the high flows will tend to be favourably selected (Liu 2020). Generally, previous decompositions of NSE (Murphy 1988; Gupta *et al.* 2009), like NSE, continue to yield a wide range of values from negative infinity to zero. This makes interpretative judgement of the model performance not straightforward. For instance, it was increasingly believed in hydrological modelling studies that model performance with positive (negative) values of KGE would be taken as good (bad). However, Knoben *et al.* (2019) recently showed that the direct use of mean flow benchmark by KGE indicates that, actually, a model's improvement starts from KGE equal to -0.41 even if the KGE values are still negative. Eventually, modellers were cautioned not to let their understanding of the ordinary NSE (Equation (1)) guide them in the interpretation of KGE (Equation (2)) (see details in Knoben *et al.* 2019). Furthermore, sampling uncertainty in the KGE estimator is substantial due to the heavy tails of the probability distribution of squared errors between model simulations and observations (Clark *et al.* 2021).

Another problem with metrics coined from previous decompositions of NSE (Murphy 1988; Gupta *et al.* 2009; Liu 2020) is that they assume linearity of the relationship between X and Y and this makes them rely on r. It is worth noting that X and Y may be highly dependent yet their dependence cannot be detected by a classical dependence measure (Székely *et al.* 2007; Chaudhuri & Hu 2019).

Preamble to the new model skill score

Suppose we want to fit a straight line to n pairs of X and Y, the equation that can be used is $Y = \alpha + \beta X$ where α is the intercept term and β denotes the least squares linear regression slope. Estimates of α and β, respectively denoted as $\hat{\alpha}$ and $\hat{\beta}$ to minimize the function $\sum (Y - (\alpha + \beta X))$ so as to guarantee the 'best fit' of the regression line, can be given by $\hat{\beta} = r(s_Y/s_X)$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$. Substituting these notations $\hat{\alpha}$ and $\hat{\beta}$ into $Y = \alpha + \beta X$ yields an equation for the fitted Y (or $\hat{Y}$) such that $\hat{Y} = \hat{\alpha} + \hat{\beta}X$. We can substitute expression for $\hat{\alpha}$ in $\hat{Y}$ and this leads to $\hat{\beta} = (\hat{Y} - \bar{Y}) \times (X - \bar{X})^{-1}$ and equating expressions for $\hat{\beta}$ from $\hat{\beta} = r(s_Y/s_X)$ and $\hat{\beta} = (\hat{Y} - \bar{Y}) \times (X - \bar{X})^{-1}$ yields

$$r = \frac{(\hat{Y} - \bar{Y})}{(X - \bar{X})} \times \frac{s_X}{s_Y} \qquad (3)$$

When we are dealing with standardized series, Equation (3) shows that r plays an important role in the regression line. Thus, so long as the appropriate model for the standardized data points is the regression line and $|r| < 1$, we find the fitted Y closer to its mean than X is to its mean. This phenomenon is referred to as regression toward the mean. Furthermore, for non-parametrically transformed series, the term r equals the linear regression slope (Onyutha 2020).

When the term α exists in a linear regression model, R^2, which can be expressed as the ratio of the explained sum of squares to total sum of squares, is actually equal to r^2 (see, e.g., Greene 1997: p. 253) such that $r = s_{XY} \times (s_X \times s_Y)^{-1}$ and $R^2 = r \times r$. We can represent the second r in $R^2 = r \times r$ using the expression in Equation (3) such that:

$$R^2 = r \times \left(\frac{s_X}{s_Y}\right) \times \left(\frac{\hat{Y} - \bar{Y}}{X - \bar{X}}\right) \qquad (4)$$

Just like for previous decomposed components of NSE as seen for instance in Equation (2), it is conspicuous that the key or common terms of R^2 from Equation (4) include $r, s_X, s_Y, \bar{X}$ and $\bar{Y}$. Furthermore, what is clear is how different these terms are arranged or combined for NSE (Equation (2)) and R^2 (Equation (4)). The difference comes about because the decomposition of NSE is via the ratio of MSE to s_X while that of R^2 is based on the ratio of s_{XY}^2 to $(s_X \times s_Y)$. In both NSE (Equation (2)) and R^2 (Equation (4)), the term r quantifies strength of linear relationship between X and Y. In Equation (4), the ratio of s_X to s_Y compares variability in X and Y. The last part of Equation (4) or the ratio of $(\hat{Y} - \bar{Y})$ to $(X - \bar{X})$ compares deviations of X and points on the regression line from $\bar{X}$ and $\bar{Y}$, respectively.

The new model metric, *E*

In this paper, the new model metric E (which for lack of an appropriate name can simply be referred to as Onyutha efficiency) is derived based on an analogy to the R^2 in Equation (4). Furthermore, E carefully considers key terms r, s_X, s_Y, $\bar{X}$ and $\bar{Y}$ as obtained for NSE (Equation (2)) and R^2 (Equation (4)). The combination of these terms r, s_X, s_Y, $\bar{X}$ and $\bar{Y}$ in the formulation of E is done in a way to ensure that E varies from zero to one while addressing existing issues of R^2 as highlighted in the introduction section. This is also done while taking into account the fact that an assumption of linearity of X and Y or the use of a classical dependence measure r as in NSE (Equation (2)) and R^2 (Equation (4)) can be violated in the case of a non-linear relationship between X and Y. Thus, instead of r as the first term of E, we use distance correlation (r_d) between X and Y (Székely *et al.* 2007), such that, $r_d = 0$ if $d_{XX} \times d_{YY} = 0$, otherwise

$$r_d = \frac{d_{XY}}{\sqrt{d_{XX} \times d_{YY}}} \tag{5}$$

where, for instance, d_{XY} denotes distance covariance of X and Y. When values of r_d are zero and one, it means X and Y are uncorrelated and perfectly correlated, respectively. Thus, r_d values of zero and one indicate poor and good model performance, respectively, with respect to correlation between X and Y. Actually, it should be pointed out that when (i) X and Y are of large n, and/or (ii) there is a large number of model simulations, the arduous computational demand in obtaining the value of r_d, especially through the use of sample distance covariance based on modified Euclidean distances, as suggested by Székely *et al.* (2007), can pose a crucial limitation for use of r_d in E. Fortunately, this limitation was substantially addressed by Chaudhuri & Hu (2019) through a fast algorithm for computing r_d. Thus, the recommended algorithm for computing r_d for use in E should be the one which is computationally not slower than that of Chaudhuri & Hu (2019).

The second component A of the metric E is to assess performance of the model with respect to variability of observations. The values of A are expected to vary from zero to one to indicate poor and good model performance, respectively. In the direct use of the ratio of s_X to s_Y as in Equation (4), the term A could be greater than one especially when s_Y is less than s_X. To ensure that the maximum value of the component A is one, we instead make use of the ratio of $min(s_X, s_Y)$ to $max(s_X, s_Y)$ where min and max denote the minimum and maximum, respectively, of any two values. This will shortly be seen important for obtaining a revised R-squared equation. For the new metric E, we can use covariance instead of standard deviation. This is in line with the fact that sample variance is basically the covariance of a variable with itself. Furthermore, since in some cases, the data may be non-normally distributed, the computation of A in terms of d_{XX} and d_{YY} becomes relevant for obtaining E given that we are also using r_d instead of r as the first term of E. Therefore, $A = 0$ if $d_{XX} = 0$ and $d_{YY} = 0$, otherwise

$$A = \frac{min(d_{XX}, d_{YY})}{max(d_{XX}, d_{YY})} \tag{6}$$

It makes intuitive sense that we do not need a model for prediction of X when $d_{XX} = 0$. Importantly, there can exist a case when the difference between $\bar{X}$ and $\bar{Y}$ (or unconditional bias) is large, indicating poor model performance, while we have an ideal performance in terms of $r_d = 1$ and $A = 1$. Thus, there is a need for the third term B to quantify model bias. To do so, we can realize that R^2 lacks capacity to quantify bias because it considers deviations of $\hat{Y}$ and X from $\bar{Y}$ and $\bar{X}$, respectively, and not a common baseline. Thus, for E we compare the squared deviations of both X and Y from a common baseline $\bar{X}$. The deviations are squared to avoid negative and positive resulting values cancelling out. Thus, the term B to quantify deviations of X and Y from a common baseline (or mean of X) can be given by $B = 0$ if $\sum (X - \bar{X})^2 = 0$ and $\sum (Y - \bar{X})^2 = 0$, otherwise

$$B = \frac{min\left(\sum (X - \bar{X})^2, \ \sum (Y - \bar{X})^2\right)}{max\left(\sum (X - \bar{X})^2, \ \sum (Y - \bar{X})^2\right)} = \frac{min(S_X^2, S_{YD}^2)}{max(S_X^2, S_{YD}^2)} \tag{7}$$

where S_X^2 is the sample variance of X and S_{YD}^2 denotes the conditional variance of Y with respect to $\bar{X}$ such that $S_{YD}^2 = (n-1)^{-1} \sum (Y - \bar{X})^2$

Finally, the metric E can be given by

$$E = r_d \times A \times B \tag{8}$$

In other words, to compute the metric E we consider r_d combined with two measures of variability of X compared with that of Y. These two measures of variability are in terms of the metrics A and B. One measure of variability is unconstrained and the other constrained. Unconstrained variability (or the term A) is obtained when we make use of standard deviations of X and Y. However, constrained variability (or the term B) is when we consider the deviations of X and Y from the mean of X as a common baseline. For bias, we make use of the difference between the average of X and that of Y. When we talk about variance and bias, there can exist a number of cases to consider including (i) high variance and large bias, (ii) high variance and small bias, (iii) low variance and large bias, and (iv) low variance and small bias. In cases (i) and (ii), we may be dealing with underfitting and overfitting, respectively. To ensure a balance between bias and variance during calibration based on objective function being introduced, the variances of both X and Y series are constrained to the mean of X (as a common baseline), and thus the term B. Finally, E is zero in any of the cases where $r_d = 0$, $A = 0$, or $B = 0$, otherwise the general formula of E is

$$E = r_d \times \frac{min(d_{XX}, d_{YY})}{max(d_{XX}, d_{YY})} \times \frac{min(S_X^2, S_{YD}^2)}{max(S_X^2, S_{YD}^2)} \tag{9}$$

Values of E in Equation (9) vary from zero to one indicating poor and ideal model performance, respectively. What should not escape quick notice is that by replacing r_d, d_{XX}, and d_{YY} of Equation (9) with absolute values of r, s_X, and s_Y, respectively, we obtain the revised R^2 version (which can be denoted as RR^2 and can be used interchangeably with RRS) whose value is taken as zero when $s_X = 0$ and $(s_X \times s_Y) = 0$, otherwise

$$RRS = |r| \times \frac{min(s_X, s_Y)}{max(s_X, s_Y)} \times \frac{min(S_X^2, S_{YD}^2)}{max(S_X^2, S_{YD}^2)} \tag{10}$$

Whereas RSS or RR^2 is simpler to apply than E and assumes linearity between X and Y, this study focused on application of E for brevity. A summary of the components of the metrics RRS and E can be seen in Figure 1. The MATLAB-based codes for computing RRS and E can be found in the Supplementary Material, Appendices A–C.

The advantages of the metric E are that it (i) allows a modeller to judge the performance of the model with respect to three measures (bias, correlation and variability), (ii) occurs over the range 0–1 and not like NSE which varies from negative infinity to one, (iii) addresses a number of the issues of the well-known R^2 and (iv) does not comprise direct squaring of model errors, something which increases the sensitivity of some metrics like NSE. These advantages also hold for the metric RRS. The main disadvantage of the metric E is that it may require automation through computer programs (like the MATLAB codes provided in Supplementary Material, Appendices A–C of this paper). Computation of the metric RRS is fast and does not necessarily require automation since it can be easily implemented, for instance in Ms Excel. The main disadvantage of RRS is that it has a component which works on the assumption that the relationship between the two given variables is linear.
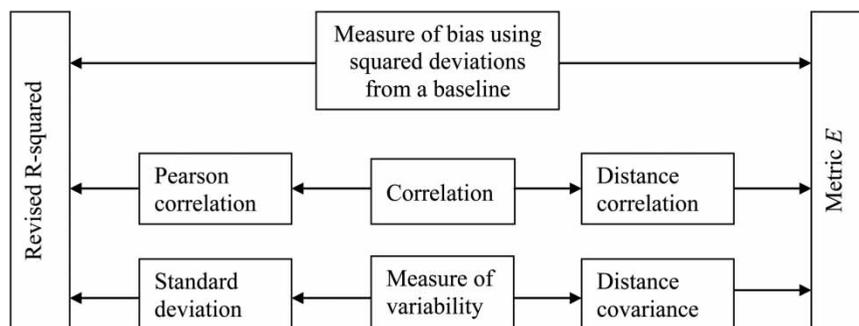

Figure 1 | Components of metrics RRS and E.

CASE STUDY

Data and selected models

It is a common practice to compare a new method with existing ones in hydrology. In this study, two catchments were selected from different climatic regions. The first catchment was of the upper Blue Nile flow observed at El Diem in Ethiopia, Africa. The catchment area of the upper Blue Nile catchment was 176,000 km^2. Quality controlled daily hydro-meteorological data comprising potential evapotranspiration (PET) and rainfall for the upper Blue Nile covering the period 1980–2000 were adopted from a previous study (Onyutha 2016). The second catchment was for the Jardine River found in North Queensland, Australia. Hydro-meteorological datasets including daily catchment runoff, catchment-wide rainfall and evapotranspiration over the Jardine River catchment were obtained from the website of 'eWater toolkit' via https://toolkit. ewater.org.au/ (September 9, 2020). The stream flow data were for gauging station no. 927001 with a catchment area of 2,500 km^2. The adopted daily PET, river flow and rainfall covering the period 1974–1989 can be found in a folder named 'Data' under Rainfall Runoff Library (RRL). To download and install RRL, one needs to first register via https://toolkit. ewater.org.au/member/CreateUser.aspx (June 9, 2020).

Two hydrological models were selected to be applied to the selected catchments. These models included Nedbør-Afstrømnings-Model (NAM) (Nielsen & Hansen 1973) and the Hydrological Model focusing on Sub-flows' Variation (HMSV) (Onyutha 2019). These models were adopted for illustration because of their lumped conceptual frameworks or structures which are compatible with the adopted catchment-wide averaged PET and rainfall. Daily PET and rainfall series were used as model inputs. The output of each model was the modelled flow and this was compared with the observed discharge.

Comparison of the new efficiency E with other 'goodness-of-fit' metrics

For calibration of HMSV and NAM, the strategy for automatically changing model parameters was based on the generalized likelihood uncertainty estimation (GLUE) framework of Beven & Binley (1992). As a Bayesian approach, GLUE requires several sets of model parameters which were randomized within stipulated limits. For each set of parameters, an objective function was selected and used to perform 10,000 sets of simulations by HMSV or NAM. The objective functions included the NSE, RSS, KGE, E, index of agreement (IOA) (Willmott 1981) and Taylor skill score (TSS) (Taylor 2001). IoA and TSS were computed using

$$IOA = 1 - \frac{\sum (X - Y)^2}{\sum (|X - \bar{X}| + |Y - \bar{X}|)^2} \tag{11}$$

$$TSS = \frac{4(1 + R_d)}{(\hat{s}_Y + 1/\hat{s}_Y)^2 (1 + R_0)} \tag{12}$$

where $\hat{s}_Y$ is the normalized s_Y and R_0 denotes the maximum correlation attainable based on another term R_d such that $R_d = n^{-1} \times \sum (X - \bar{X})(Y - \bar{Y}) \times (s_X s_Y)^{-1}$. To determine R_0 to be used in Equation (12), NAM and HMSV were each simulated 6,000 times using GLUE strategy. During each simulation run, R_d was computed. The maximum of the 12,000 R_d values (6,000 from each model) was taken as R_0.

Using each of the objective functions (including NSE, RSS, E, IOA, KGE, and TSS), both HMSV and NAM were calibrated based on the GLUE strategy. There was a total of 10,000 calibration runs with each objective function. During calibration using a particular objective function, values for other objective functions were also calculated. In other words, at the end of the calibration, each of the objective functions including NSE, RSS, E, IOA, KGE and TSS had 10,000 values. Values of one objective function were plotted against those for other metrics. The optimal parameters were those in the set which yielded the best (or highest) value of the objective functions used in this study. Modelled runoff series obtained based on the objective functions (NSE, KGE, E, TSS, IOA and RRS) were also graphically compared. Furthermore comparison was made on (i) R-squared and RRS, (ii) r and r_d, as well as (iii) RRS and E.

RESULTS AND DISCUSSION

Comparison of E, revised R-squared, Pearson correlation and distance correlation

Figure 2 shows comparison of the various relevant 'goodness-of-fit' metrics. The scatter points in the plots of r versus r_d do not fall along the bisector or diagonal dashed line (Figure 2(a) and 2(b)). Values of r were mostly greater than those of r_d. Thus,
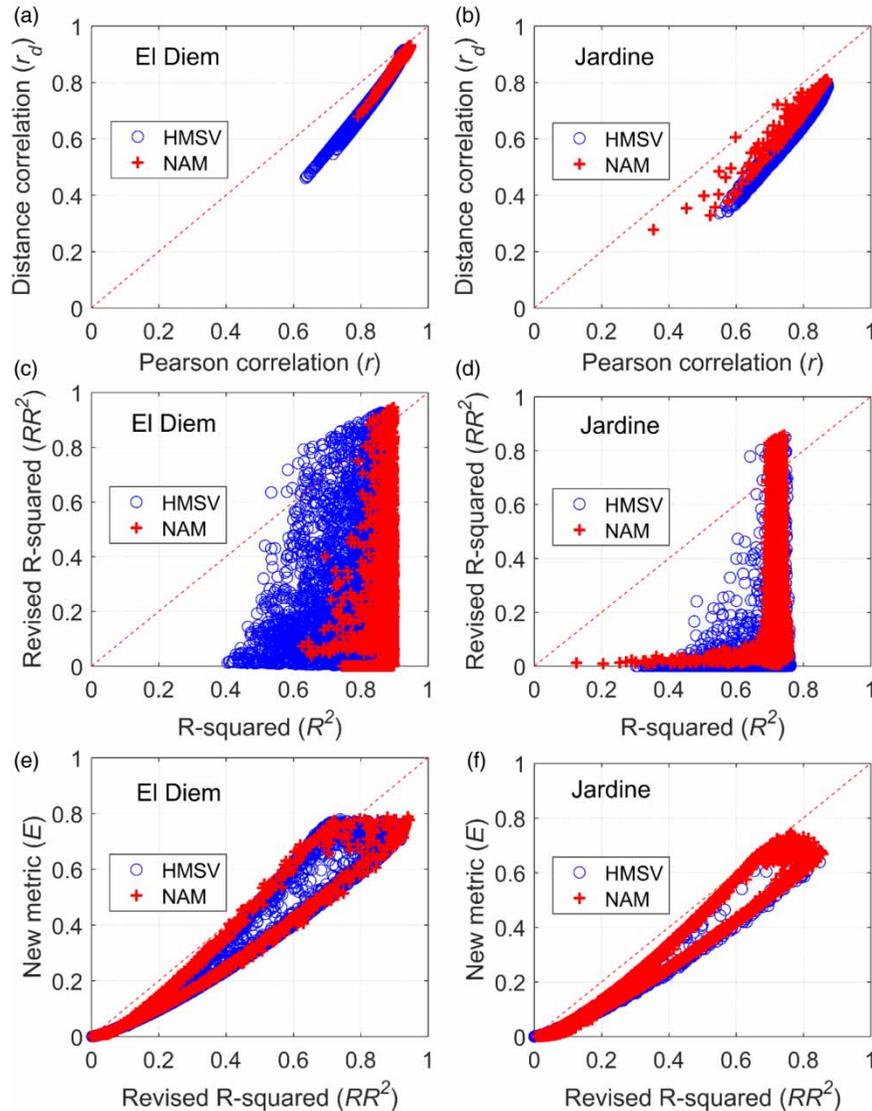
Figure 2 | Comparison of (a) and (b) r and r_d, (c) and (d) R^2 and RR^2, and (e) and (f) RR^2 and E.

when the relationship between X and Y is not linear, the use of r becomes inadequate in evaluating the model's 'goodness-of-fit'. Values of the original R^2 can be seen to be greater than the revised one, RR^2, for almost all the cases (Figure 2(c) and 2(d)). The original R^2 yields values close to its maximum even when there are large mismatches between X and Y. Results shown in Figure 2(c) and 2(d) indicate that RR^2 is far better than R^2 in evaluating model performance or assessing predictive power of a model. This is because RR^2 combines measures of bias, variability and correlation. Nevertheless, the scatter points in the plots of RR^2 versus E fall mostly below the bisector. Thus, RR^2 yields values which, in most cases, are greater than the new metric E. The differences between E and RR^2, as reflected in Figure 2(e) and 2(f), are because (i) E uses d_{XX} and d_{YY} while RR^2 makes use of s_X and s_Y, and (ii) E comprises r_d while RR^2 applies r. The difference between E and RR^2 shows the limitation of the assumption of linear relationships between X and Y while evaluating model performance. Conclusively, the use of RR^2 should be after confirming the linearity of the relationship between X and Y. This could be, for instance, through a simple scatter plot of X versus Y.

Values of *E* components based on calibrations of HMSV and NAM

Figure 3 shows results of model performance with respect to variability (A), bias (B) and distance correlation (r_d). It can be noted that r_d was, in most cases, greater than values of both A and B. Four points P, Q, R and S are marked on Figure 3 for
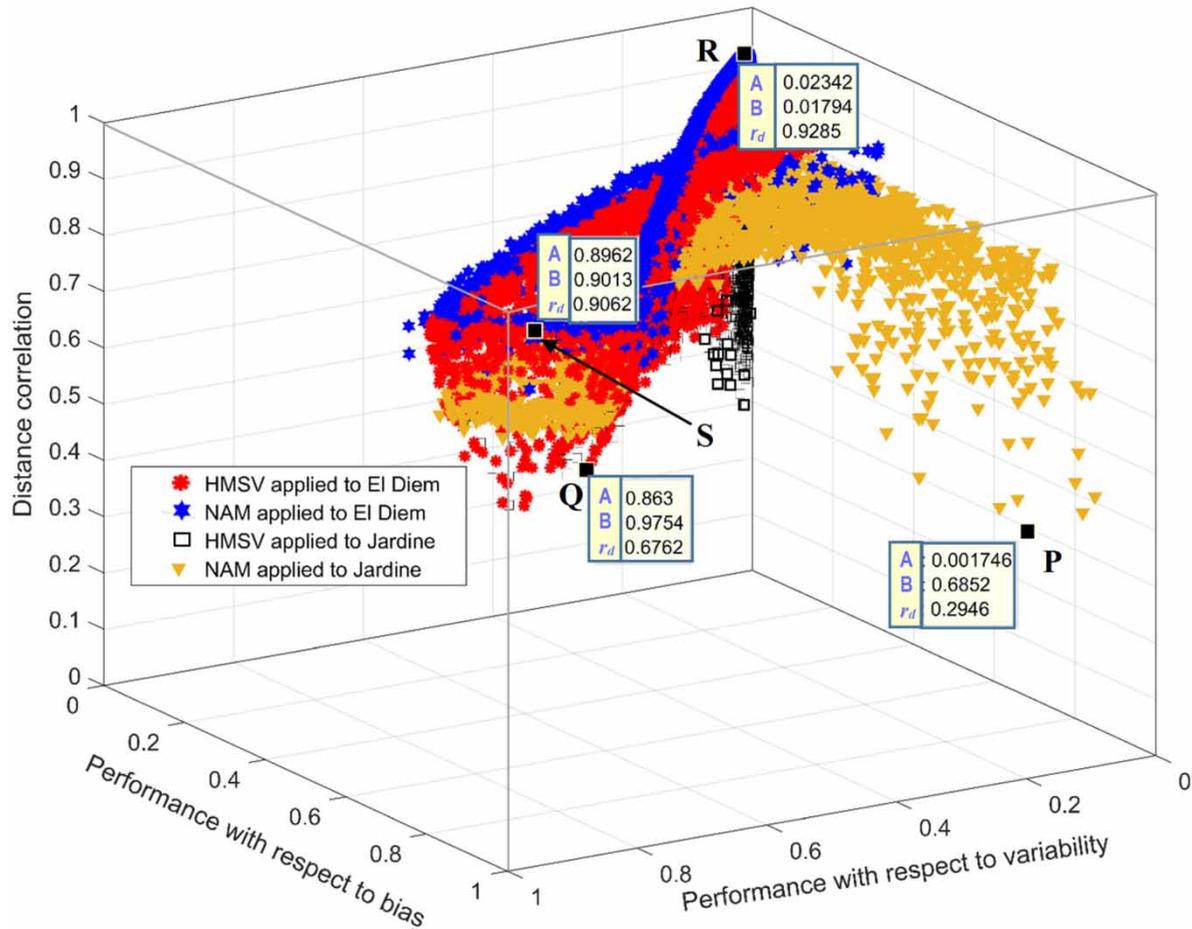
Figure 3 | Comparison of model performance in terms of the various components of E.

illustration to highlight what a modeller misses out by using other 'goodness-of-fit' metrics such as NSE, IOA, TSS and RSS. Values at point P show that the model performance was fair with respect to bias ($B = 0.6852$) but poor regarding both variability ($A = 0.0017$) and distance correlation ($r_d = 0.2946$). At point R, the model is good in performance with respect to only ($r_d = 0.9285$). At point Q, the model performed better regarding variability ($A = 0.8630$) and bias ($B = 0.9754$) than distance correlation ($r_d = 0.6762$). The best model performance for the selected illustrative points was at S with E equal to 0.7320. Even at point S, the model performed slightly better with respect to bias ($B = 0.9013$) and distance correlation ($r_d = 0.9062$) than variability ($A = 0.8962$).

The illustrations in Figure 3 show that a model's performance can be poor because of its (i) large bias, (ii) limited capacity to reproduce variability in observed data, or (iii) reduced capability to generate outputs which can resonate well with observed data. Given these known reasons, a modeller can focus accordingly on which parameter is relevant for adjustments to improve the model performance during calibration. To do so, one needs to have an expert judgement of the system under consideration. For instance, if we are dealing with a hydrological system, one reason why a model can largely be biased is reduced model capacity due to flawed modelling concepts that are unable to capture impacts of human activities on hydrology. To exemplify this, we need to think about a river whose flow is abstracted at various locations along its channel length for irrigation, industrial use and water supply to towns or cities. In another case, there can be some return flows into the river, for instance, in the form of treated industrial effluents, and discharge from an irrigation field. In such cases, the recorded discharge in the river can be different from the actual one. Eventually, there can exist a systematic difference between the simulated and measured river flows. Some models (like HMSV) give provisions to take into account possible flow returns or abstractions from a catchment. Specifically, the HMSV caters for flow returns or abstractions using a value (with the unit as of flow being modelled) that needs to be entered by the modeller. In case the model performs poorly with respect to variability or distance correlation, some

parameters to be adjusted are those which take care of recession of overland flow, inter flow and base flow. Such parameters, for instance, in NAM include time constant for interflow (CK_{IF}, day), time constant for base flow (CK_{BF}, day), overland flow runoff coefficient (CQ_{OF}). In the same line, HMSV makes use of base flow recession constant (t_1, day), interflow recession constant (t_b, day) and overland flow recession constant (t_u, day) to control variability in flow.

## Comparing the new metric *E* with other objective functions

Figure 4 shows comparison of various objective functions (NSE, RSS, E, IOA, KGE and TSS). For negative values of NSE, KGE and RSS, *E* is zero or nearly so (Figure 4(a)–4(c)). When *E* is zero, TSS and IOA can attain values as high as 0.6 and 0.7,
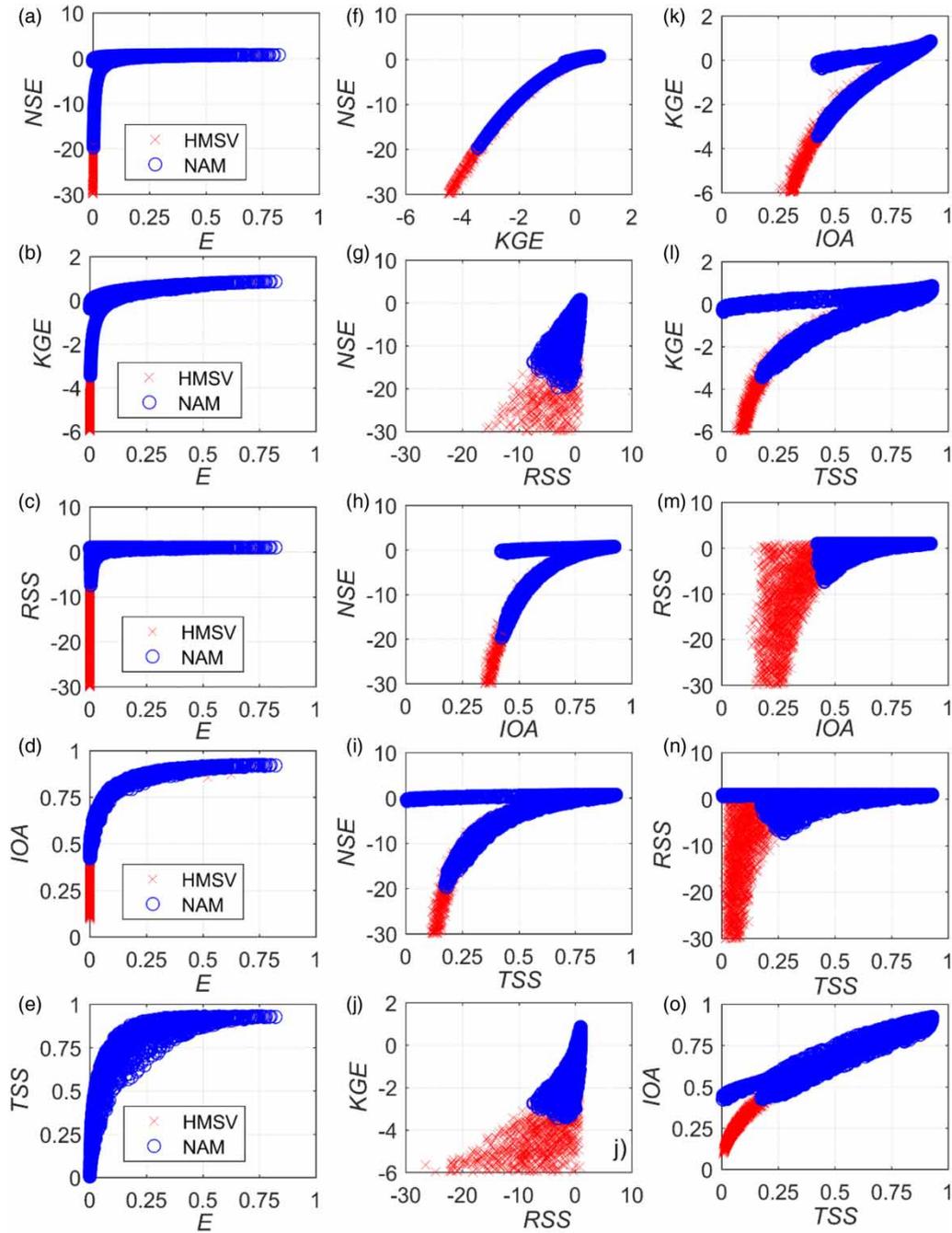
Figure 4 | Plots of (a)–(e) *E* against NSE, KGE, RSS and TSS; (b)–(i) NSE versus KGE, RSS, IOA and TSS; (j)–(l) KGE against RSS, IOA and TSS; (m) and (n) RSS versus IOA and TSS; and (o) IOA versus TSS.

respectively (Figure 4(d) and 4(e)). It can be noted that RSS, IOA and TSS increase more rapidly towards the maximum value of one than E (Figure 4(c) and 4(e)). NSE was shown to be more negative than KGE, especially for values less than zero (Figure 4(f)). KGE is a variant of NSE and this is why scatter points in the plot of KGE versus NSE depict a polynomial (like nearly linear) relationship (Figure 4(f)). Eventually, the relationship between NSE and IOA is comparable with that of KGE and IOA (Figure 4(h) and 4(k)). Similarly, scatter points in the plots of NSE versus TSS (Figure 4(i)) are to a large extent comparable with that of KGE versus TSS (Figure 4(l)). Furthermore, RSS versus NSE (Figure 4(g)) is also comparable with that of RSS against KGE (Figure 4(j)). For very low (even negative) NSE, KGE and RSS, both IOA and TSS yielded large values (Figure 4(g)–4(i), 4(k)–4(n)). IOA has a major drawback of giving high values (close to 1) even for a poorly fitted model (Krause *et al.* 2005). To address the problems related to the use of IOA, Willmott *et al.* (2012) reformulated the IOA with respect to the bounds of the values. In a relevant development, Legates & McCabe (2013) remarked that the refinement made by Willmott *et al.* (2012) to extend the bound of the original IOA such that it starts from -1 to 0 was unnecessary. Other limitations of the refined IOA can be found elaborately given by Legates & McCabe (2013). Generally, IOA is more comparable with TSS than other metrics (Figure 4(o)). In other words, scatter points of TSS and IOA are nearly
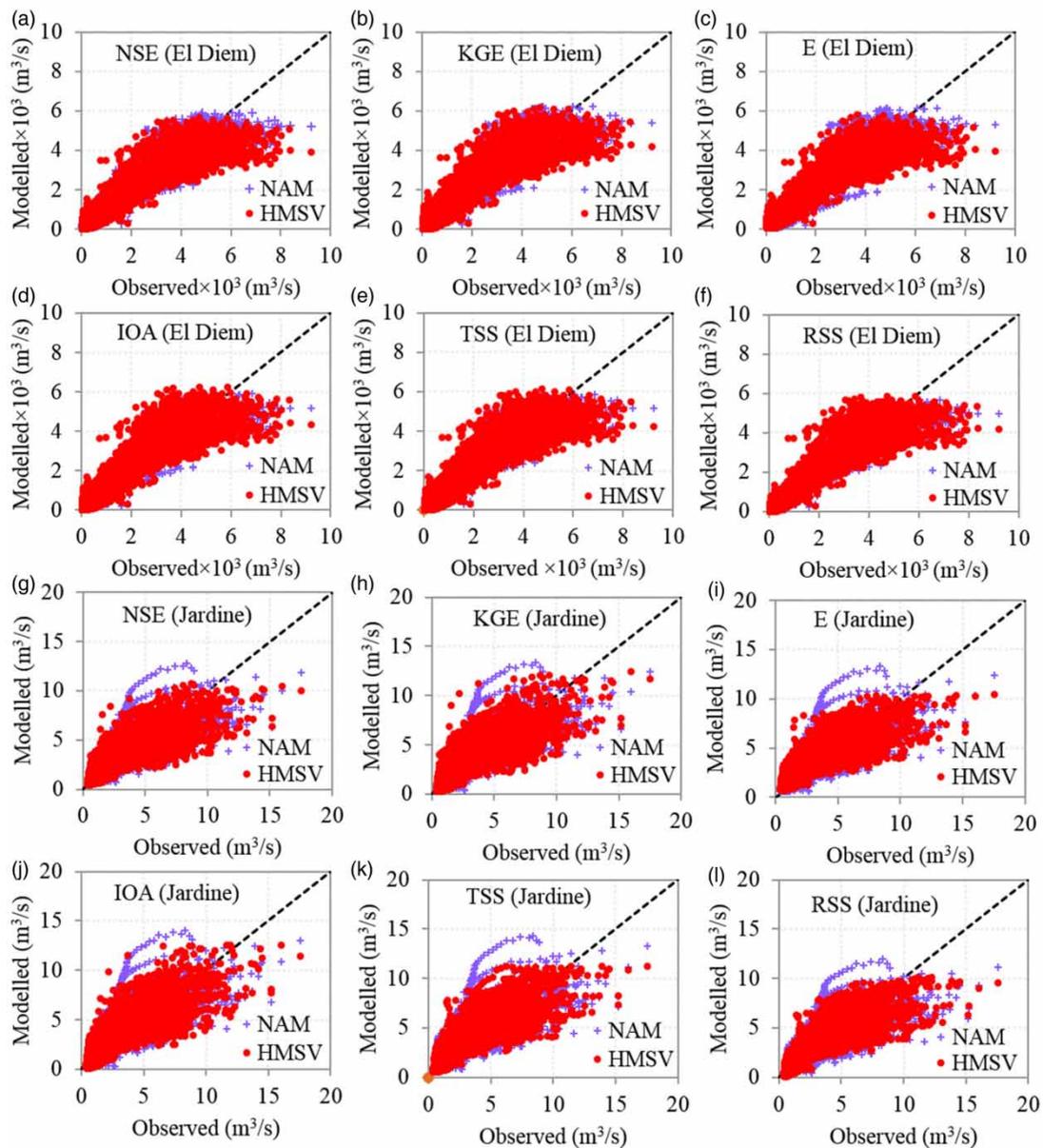
Figure 5 | Observed versus modelled flow based on HMSV and NAM applied to (a)–(f) El Diem and (g)–(l) Jardine catchments.

linear (Figure 4(o)). This explains why plots of E versus IOA (Figure 4(d)) are comparable with that of E versus TSS (Figure 4(e)). Results in Figure 4(d) generally show the acceptability of the new metric E in comparison with other existing 'goodness-of-fit' measures.

Figure 5 shows plots of observed and modelled flows obtained based on various objective functions. Generally, results from the HMSV agree with those of NAM. However, because HMSV and NAM have different structures and parameters, some slight differences between results of the two models can be seen in terms of the spread of scatter points around the bisector. The idea behind comparisons in Figure 5 was not to show which objective function leads to the best model outputs but to show the comparability of the results. For an ideal model, the scatter points would fall along the dashed diagonal line (also called the bisector). The spread of the scatter points around the bisector indicates model errors. For all the models and objective functions, some extreme events were overestimated or underestimated. Results for the various objective functions are comparable as expected when the models were applied to both El Diem and Jardine catchments (Figure 5(a)–5(l)). This shows the acceptability of the new metric E as an objective function for calibrating models.

CONCLUSIONS

Coefficient of determination (R^2) is arguably the most widely applied 'goodness-of-fit' measure in modelling or prediction of hydrological, ecological, agricultural and climatic categories of the environmental systems. However, there are a number of issues which are well-known in the application of R^2 including the fact that it: (i) does not quantify model bias; (ii) can be low for an accurate model; (iii) can be high for an imperfect model; and (iv) yields the same value when we regress observed (X) and modelled (Y) and vice versa. In fact, issue (iv) invalidates the use of R^2 as the coefficient of determination to indicate the amount of the total variance in observations explained by the model. Another commonly applied version of R^2 or the well-known Nash–Sutcliffe efficiency (NSE) (Nash & Sutcliffe 1970) is also known to have the problems of varying from negative infinity to one, and being oversensitive to extreme or large values stemming from the influence of squaring the error term. Another commonly used metric, KGE, also varies from negative infinity to one like the NSE. A model's improvement starts from KGE equal to -0.41 even if the KGE values are still negative, and eventually modellers were cautioned not to let their understanding of the ordinary NSE guide them in the interpretation of KGE values (Knoben *et al.* 2019). Furthermore, both NSE and KGE were shown to have substantial sampling uncertainties due to the heavy tails of the probability distribution of squared errors between model simulations and observations (Clark *et al.* 2021). Nevertheless, attempts to address the issues of R^2 were central to this study. Eventually, this paper (i) revisited R^2 to become RR^2 (or RRS) and (ii) introduced a new model skill score also called Onyutha Efficiency E. Both E and RRS make use of correlation, model performance with respect to variability (A) and bias (B). The differences between E and RRS lie in the forms of correlation and the term A used for each metric. The RRS is a product of (i) Pearson's correlation, (ii) a measure which compares standard deviations of X and Y and (iii) the term B. For the metric E, the term A considers distance covariances of X and Y. Results of simulations demonstrated superiority of the RRS over the original version of R-squared. By applying the metric E and RRS, the modeller can diagnostically identify and expose the systematic issues behind model optimizations based on other 'goodness-of-fit' measures such as NSE, R^2 and mean squared error.

To apply E and RR^2, the reader can find MATLAB codes provided in Supplementary Material, Appendices A–C, as well as via https://www.researchgate.net/publication/356420464 and https://sites.google.com/site/conyutha/tools-to-download (accessed: 11/21/2021).

CONFLICT OF INTEREST

The author declares no conflict of interest and no competing financial interests.

AUTHOR CONTRIBUTION STATEMENT

The entire work in this paper was based on the effort of the sole author.

CODE/DATA AVAILABILITY

The MATLAB codes to implement the new methods are included in Supplementary Material, Appendices A–C, as well as via https://www.researchgate.net/publication/356420464 and https://sites.google.com/site/conyutha/tools-to-download (accessed: 11/21/2021).

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Alexander, D. L. J., Tropsha, A. & Winkler, D. A. 2015 Beware of R^2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling* **55**, 1316–1322.

Bai, Z., Wu, Y., Ma, D. & Xu, Y.-P. 2021 A new fractal-theory-based criterion for hydrological model calibration. *Hydrology and Earth System Sciences* **25**, 3675–3690.

Barber, C., Lamontagne, J. R. & Vogel, R. M. 2019 Improved estimators of correlation and R^2 for skewed hydrologic data. *Hydrological Sciences Journal* **65**, 87–101.

Bardsley, W. E. 2013 A goodness of fit measure related to r^2 for model performance assessment. *Hydrological Processes* **27** (3), 2851–2856.

Beven, K. J. & Binley, A. M. 1992 The future role of distributed models: model calibration and predictive uncertainty. *Hydrological Processes* **6**, 279–298.

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H. G., Sivapalan, M., Stumpp, C., Toth, E. & Volpi, E. 2019 Twenty-three unsolved problems in hydrology (UPH) – a community perspective. *Hydrological Sciences Journal* **64** (10), 1141–1158.

Cameron, S. 1993 Why is the R squared adjusted reported? *Journal of Quantitative Economics* **9**, 183–186.

Chaudhuri, A. & Hu, W. 2019 A fast algorithm for computing distance correlation. *Computational Statistics & Data Analysis* **135**, 15–24.

Cheng, C.-L., Shalabh & Garg, G. 2014 Coefficient of determination for multiple measurement error models. *Journal of Multivariate Analysis* **126**, 137–152.

Chicco, D., Warrens, M. J. & Jurman, G. 2021 The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE. *MSE and RMSE in Regression Analysis Evaluation. PeerJ Computer Science* **7**, e623.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R. & Papalexiou, S. M. 2021 The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research* **57**, e2020WR029001. https://doi.org/10.1029/2020WR029001.

Ferreira, P. M. L., Paz, A. R. & Bravo, J. M. 2020 Objective functions used as performance metrics for hydrological models: state-of-the-art and critical analysis. *Brazilian Journal of Water Resources* **25** (e42), 1–15.

Garrick, M., Cunnane, C. & Nash, J. E. 1978 A criterion of efficiency for rainfall-runoff models. *Journal of Hydrology* **36**, 375–381.

Goldberg, A. S. 1991 *A Course in Econometrics*. Harvard University Press, Cambridge, MA, USA.

Greene, W. 1997 *Econometric Analysis*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ, USA.

Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. 2009 Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modeling. *Journal of Hydrology* **377**, 80–91.

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J. & Ames, D. P. 2019 Introductory overview: error metrics for hydrologic modelling-a review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling and Software* **119**, 32–48.

Jenkins, D. G. & Quintana-Ascencio, P. F. 2020 A solution to minimum sample size for regressions. *PLoS ONE* **15** (2), e0229345. https://doi.org/10.1371/journal.pone.0229345.

Jie, M. X., Chen, H., Xu, C. Y., Zeng, Q. & Tao, X. E. 2016 A comparative study of different functions to improve the flood forecasting accuracy. *Hydrology Research* **47**, 718–735.

Knoben, J. M. W., Freer, J. E. & Woods, R. A. 2019 Technical note: inherent benchmark or not? comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences* **23**, 4323–4331.

Krause, P., Boyle, D. P. & Bäse, F. 2005 Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences* **5**, 89–97.

Kvålseth, T. O. 1985 Cautionary note about R^2. *The American Statistician* **39** (4), 279–285.

Lamontagne, J. R., Barber, C. A. & Vogel, R. M. 2020 Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Research* **56** (9), e2020WR027101. https://doi.org/10.1029/2020WR027101.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**, 233–241.

Legates, D. R. & McCabe, G. J. 2013 A refined index of model performance: a rejoinder. *International Journal of Climatology* **33**, 1053–1056.

Lin, F., Chen, X. & Yao, H. 2017 Evaluating the use of Nash-Sutcliffe efficiency coefficient in goodness-of-fit measures for daily runoff simulation with SWAT. *Journal of Hydrologic Engineering* **22** (11), 05017023. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001580.

Liu, D. 2020 A rational performance criterion for hydrological model. *Journal of Hydrology* **590**, 125488.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V. & Kumar, R. 2019 On the choice of calibration metrics for 'high-flow' estimation using hydrologic models. *Hydrology and Earth System Sciences* **23**, 2601–2614.

Moriasi, D. N., Gitau, M. W., Pai, N. & Daggupati, P. 2015 Hydrologic and water quality models: performance measures and evaluation criteria. *Transactions of the ASABE* **58** (6), 1763–1785.

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

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10**, 282–290.

Nielsen, S. A. & Hansen, E. 1973 Numerical simulation of the rainfall-runoff process on a daily basis. *Nordic Hydrology* **4** (3), 171–190.

Onyutha, C. 2016 Influence of hydrological model selection on simulation of moderate and extreme flow events: a case study of the Blue Nile basin. *Advances in Meteorology* **2016** (Article ID 7148326), 1–28. https://doi.org/10.1155/2016/7148326.

Onyutha, C. 2019 Hydrological model supported by a step-wise calibration against sub-flows and validation of extreme flow events. *Water* **11** (2), 244. https://doi.org/10.3390/w11020244.

Onyutha, C. 2020 From R-squared to coefficient of model accuracy for assessing 'goodness-of-fits'. *Geoscientific Model Development Discussion*. https://doi.org/10.5194/gmd-2020-51.

Pool, S., Vis, M. & Seibert, J. 2018 Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal* **63** (13–14), 1941–1953.

Quinino, R. C., Reis, E. A. & Bessegato, L. F. 2013 Using the coefficient of determination R^2 to test the significance of multiple linear regression. *Teaching Statistics* **35**, 84–88.

Rose, S. & McGuire, T. G. 2019 Limitations of p-values and Rsquared for stepwise regression building: a fairness demonstration in health policy risk adjustment. *The American Statistician* **73**, 152–156.

Schober, P. M. D., Christa, B. & Lothar, A. S. 2018 Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia* **126**, 1763–1768.

Stoffel, M. A., Nakagawa, S. & Schielzeth, H. 2021 Partr2: partitioning R2 in generalized linear mixed models. *PeerJ* **9**, e11414. http://doi.org/10.7717/peerj.11414.

Székely, G. J., Rizzo, M. L. & Bakirov, N. K. 2007 Measuring and testing independence by correlation of distances. *The Annals of Statistics* **35** (6), 2769–2794.

Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research* **106**, 7183–7192.

Willmott, C. J. 1981 On the validation of models. *Physical Geography* **2**, 184–194.

Willmott, C. J., Robeson, S. M. & Matsuura, K. 2012 A refined index of model performance. *International Journal of Climatolology* **32**, 2088–2094.

Yates, J. F. 1982 External correspondence: decomposition of the mean probability score. *Organizational Behavior and Human Performance* **30**, 132–156.

Ye, L., Gu, X., Wang, D. & Vogel, R. 2021 An unbiased estimator of coefficient of variation of streamflow. *Journal of Hydrology* **594**, 125954. https://doi.org/10.1016/j.jhydrol.2021.125954.