Environmental Modelling & Software 73 (2015) 167–174

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Assessment of three dimensionless measures of model performance

Cort J. Willmott [a, *], Scott M. Robeson [b], Kenji Matsuura [a], Darren L. Ficklin [b]

[a] Department of Geography, University of Delaware, Newark, DE 19716, USA
[b] Department of Geography, Indiana University, Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 18 June 2015
Received in revised form
11 August 2015
Accepted 12 August 2015
Available online xxx

Keywords:
Dimensionless accuracy indices
Model-performance statistics
Streamflow model evaluation

ABSTRACT

Pertinent characteristics of three dimensionless and comparable model-performance or model-efficiency measures are examined and compared. Our model-assessment recommendations apply to many types of environmental models. Representing measures based on sums-of-squared errors or "quadratic measures" (by far, the most widely used class) is Nash and Sutcliffe's (1970) well-known efficiency measure (E). Our assessments of E, by and large, also apply to similar-in-form benchmark measures (Seibert, 2001). Legates and McCabe's (1999) version of E (E_1) and Willmott et al.'s (2012) refined index of agreement (d_r) represent the less-often-employed but more interpretable class of measures based on sums of error magnitudes. Conceptual and algebraic arguments are used in conjunction with assessments of many parameter sets of the Soil and Water Assessment Tool (SWAT) hydrologic model, over the period 1950–2005. Our findings suggest that, of the three measures, d_r has the broadest utility, followed in order by E_1 and E.

1. Introduction

Process-based numerical models are developed and applied within nearly all of the environmental sciences and engineering, although here we draw mainly from work in hydrology and climate science (e.g., Singh and Frevert, 2006; Ficklin et al. 2012; Bring and Destouni, 2014; Ashofteh et al. 2015). An essential aspect of these endeavors is estimating and describing how well the models perform; or, in slightly different terms, how well model-estimated variables compare with corresponding, reliable observations (cf, Willmott et al. 1985; Bennett et al. 2013). Approaches vary from traditional quantitative comparisons of statistical or numerical criteria and associated graphics (e.g., Willmott, 1981, 1982; Willmott et al. 1985; Gupta et al. 2009) to expert assessments of similarity among model-estimated and observed variables (e.g., Crochemore et al. 2015). Our interest here, however, is on efficacious statistical or numerical comparisons of model-produced estimates or predictions with corresponding, independent (of the model) and reliable observations. This remains "an open and critical issue" (Berthet et al., 2010) because of nontrivial questions about sums-of-squared-based or "quadratic" error measures. More specifically, our focus here is on three increasingly used dimensionless measures of accuracy that, we believe, comprise a key subset of the wider array of approaches to assessing model performance (Bennett et al. 2013).

Important but often underappreciated distinctions between model-performance and model-training statistics (cf, Willmott, 1981, 1982; Gupta et al. 2009; Ehret and Zehe, 2011) are briefly summarized here, as understanding of several of our key points depends on them. Most model-training statistics emerged from well-known statistical techniques for fitting functions to data (Draper and Smith, 1998), primarily by minimizing sums-of-squares deviations between corresponding sets of model-produced estimates and observations. The mathematical and statistical properties of these approaches are well known (Draper and Smith, 1998) and familiar to most environmental scientists and engineers (e.g., Gupta et al. 2009; Bennett et al. 2013). Model-training statistics have been useful in estimating model-parameter sets, in recalibrating fit-model parameters, and in decomposing sets of model deviations from observations (cf, Willmott, 1981, 1982; Gupta et al. 2009). There are, however, frequently overlooked but important interpretational problems associated with the application of these sums-of-squares-based statistics to the evaluation of model performance. These problems arise because sums-of-squared-based statistics—like the frequently reported root-mean-square error—are influenced by three independent variables: the mean of the individual error

* Corresponding author. Department of Geography, University of Delaware, Newark, DE 19716, USA.
E-mail addresses: willmott@udel.edu (C.J. Willmott), srobeson@indiana.edu (S.M. Robeson).

http://dx.doi.org/10.1016/j.envsoft.2015.08.012
1364-8152/

magnitudes, the variability among the error magnitudes, and the number of observations or domain of integration (Willmott and Matsuura, 2005, 2006). Despite common usage (e.g., Gupta et al. 2009; Bennett et al. 2013; Ashofteh et al. 2015), the mix of influences from these three variables on a sum-of-squares-based-error statistic makes meaningful interpretations of it or comparisons among two or more such statistics difficult or even impossible. Many routinely reported sums-of-squared-errors statistics are (virtually always) overestimates of average error-magnitude (Willmott and Matsuura, 2005, 2006).

Unlike in model training, when evaluating model performance, the model structure and its parameters are (or should be) already known. As a result, there is no need to solve for them by minimizing sums-of-squared-error or -deviation measures. Further, the observations to which the model-produced estimates are being compared should be independent of the model. The primary value of model-performance-error statistics should be in what they convey about the size and nature of model-prediction/estimation error and not in parameter estimation, calibration or decomposition. Several model-performance-error statistics that are based on sums of absolute values (magnitudes) of errors, on the other hand, have straightforward interpretations and can be readily compared; so, they generally are preferable. Our assessments and comparisons of dimensionless model-performance measures are informed by these considerations.

While there is a wide variety of recommended model-assessment or -efficiency statistics available (cf, Nash and Sutcliffe, 1970; Legates and McCabe, 1999; Seibert, 2001; Gupta et al. 2009; Willmott, 1981, 1982; Willmott et al., 2012), at a minimum, we encourage researchers to evaluate, report, and interpret at least one dimensioned and one dimensionless measure of average error-magnitude. The mean bias error (MBE) also can be useful in that it describes average over- or under-prediction, although we do not examine it here. Our judgment (Willmott and Matsuura, 2005, 2006) is that the mean absolute error (MAE) is the preferred measure of average error-magnitude and that one of a handful of available dimensionless indices also should be reported and assessed (Willmott et al., 2012). Oyler et al.'s (2015) application of the MAE and Willmott et al.'s (2012) refined index of agreement illustrate the usefulness of this approach. Inspection of one-to-one scatterplots of predicted on observed variables and related graphics also are essential and can reveal distributional aspects of the relationship, systematic biases, and outliers (Bennett et al. 2013). Our focus within this paper is on comparing the available dimensionless indices of model performance for environmental (especially hydrologic) applications.

In a "Short Communication", Willmott et al. (2012) compared six of these dimensionless indices of model performance, based on conceptual and algebraic arguments made in conjunction with sensitivity assessments using a pseudo-random number generator. Within this paper, we build on Willmott et al.'s analyses and compare, in greater depth, the two most commonly used in the hydrologic sciences and engineering—Nash and Sutcliffe's (1970) coefficient of Efficiency (E) and Legates and McCabe's (1999) refinement of E (E_1)—with the refined index of agreement (d_r) offered by Willmott et al. (2012). Refinements and decompositions of Nash and Sutcliffe's E have appeared in the literature many times (e.g., Legates and McCabe, 1999; Seibert, 2001; Gupta et al. 2009) but, since the original measure remains representative and is still widely employed, it serves well here to characterize the class of sums-of-squares-based error measures. The behavior of E, for example, parallels and, in turn, represents the behavior of benchmark–based measures discussed by Seibert (2001). Our comparisons here emphasize both the salient properties of these three measures and how well each one illuminates the performances of alternate versions (parameterizations) of the Soil and Water Assessment Tool (SWAT) streamflow model (Arnold et al. 1998). Our discussion begins with necessary definitions and algebraic and conceptual comparisons, followed by performance assessments of an array of alternate parameterizations of the SWAT streamflow model, applied to predicting flow within the American, Sacramento and Kern River basins in California (Fig. 1).

2. Background

For many years, Nash and Sutcliffe's E and refinements of it (e.g., Seibert, 2001; Gupta et al. 2009) have been widely applied in the hydrologic sciences and engineering but Legates and McCabe's important modification of E (E_1), in particular, is becoming increasingly popular, according to citation counts. Willmott et al.'s revised index (d_r) is newer but already it too has found a number of applications (e.g., Gaitan et al. 2014; Ward et al. 2013; Bring and Destouni, 2014; Oyler et al. 2015), and we think that it has considerable promise within the environmental science and engineering communities. Although the algebraic details of E, E_1 and d_r were detailed by Willmott et al. (2012), they can be rewritten and

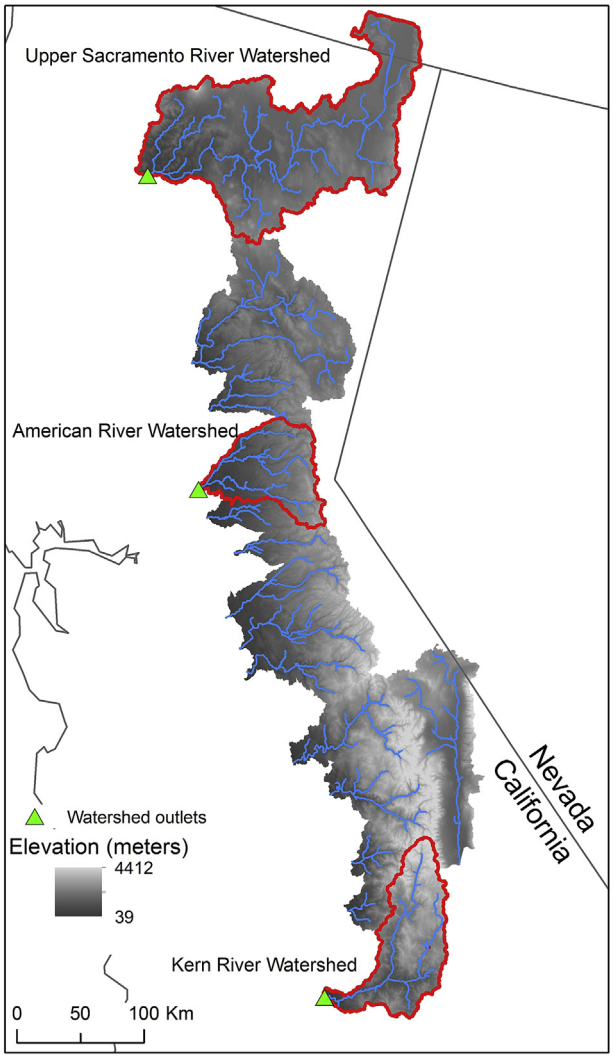

Fig. 1. American, Kern, and Upper Sacramento River watershed-model domains (outlined in red). Observed and simulated streamflow-discharge values are assessed at the watershed outlets (green triangles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

discussed more succinctly in terms of their summed and averaged components: these include; the mean-square error (MSE), the mean-absolute error (MAE), and the observed variance (s_o^2) and mean-absolute deviation (MAD) associated with the observed variable.

The averaged sums—the MSE, the MAE and the MAD—can be expressed as

$$\begin{aligned} \mathrm{MSE} &= \sum_{i=1}^{n} w_i (P_i - O_i)^2 \Big/ \sum_{i=1}^{n} w_i, \\ \mathrm{MAE} &= \sum_{i=1}^{n} w_i |P_i - O_i| \Big/ \sum_{i=1}^{n} w_i, \quad \text{and} \\ \mathrm{MAD} &= \sum_{i=1}^{n} w_i |O_i - \overline{O}| \Big/ \sum_{i=1}^{n} w_i \end{aligned} \tag{1}$$

wherein model-derived estimates or predictions (P_i; $i = 1,2,\ldots,n$) are compared with pairwise-matched observations (O_i; $i = 1,2,\ldots,n$) that are judged to be reliable, and w_i is a scaling assigned to each $(P_i - O_i)^2$, $|P_i - O_i|$ and $|O_i - \overline{O}|$ according to its hypothesized influence on the total error or total deviation (Willmott and Matsuura, 2006). For simplicity here, we let $w_i = 1.0$ for all i; thusly here, they sum to n—the number of pairs of model estimates and observations. The units of P and O are the same.

In terms of these error and deviation averages, E and E_1 can be written as

$$E = 1 - \frac{\mathrm{MSE}}{c^2 \cdot s_o^2} \quad \text{and} \quad E_1 = 1 - \frac{\mathrm{MAE}}{c \cdot \mathrm{MAD}} \tag{2}$$

with the scaling coefficient $c = 1.0$. Nash and Sutcliffe (1970) and Legates and McCabe (1999) did not explicitly include a scaling coefficient within their presentations of E and E_1; but, implicitly, their indices are scaled with $c = 1.0$ (the reason for this distinction will become apparent below). Note that for our purposes here, we compute the observed variance (s_o^2) with n rather than with another degrees-of-freedom estimate. Our d_r is

$$d_r = \begin{cases} 1 - \dfrac{\mathrm{MAE}}{c \cdot \mathrm{MAD}}, & \text{when } \mathrm{MAE} \le c \cdot \mathrm{MAD} \\ \dfrac{c \cdot \mathrm{MAD}}{\mathrm{MAE}} - 1, & \text{when } \mathrm{MAE} > c \cdot \mathrm{MAD} \end{cases} \tag{3}$$

with $c = 2.0$. The rationale for $c = 2.0$ is described below.

While all three statistics are derived from average-error (the MSE or the MAE) and average-deviation (s_o^2 or MAD) measures, each translates and conveys the relationships between the MSE or MAE and s_o^2 or MAD somewhat differently.

3. Interpretations of E, E_1 and d_r

Responses of E, E_1, and d_r to differing patterns of error size and variability are similar, but there are important differences. Perhaps the most important difference is that the ranges of E and E_1 are $(-\infty < E \le 1.0)$ and $(-\infty < E_1 \le 1.0)$, while d_r has finite lower and upper bounds $(-1.0 \le d_r \le 1.0)$. Increases and decreases in d_r and E_1 are monotonic over their entire response domains (see Fig. 3 in Willmott et al., 2012) but this monotonicity does not hold between E and E_1 [even though their algebraic structures are quite similar, see Equation (2)] or between E and d_r, since it is the square of each term that influences E rather than the absolute value of each term that influences E_1 and d_r. It should be noted that this monotonicity between changes in E_1 and d_r does not mean that E_1 and d_r are the same measure, no more than Kelvin, Celsius, and Fahrenheit are the same temperature scale. While salient aspects of the similarities and differences among E, E_1, and d_r are our central theme, our discussion emphasizes differences between E_1 and d_r because—even though these differences may appear subtle—they can be important.

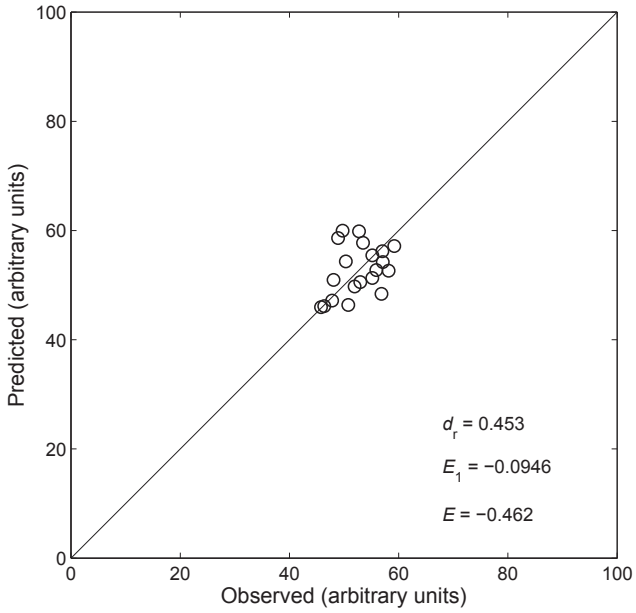

Fig. 2. Twenty randomly selected hypothetical observations and corresponding hypothetical model-predicted values. The hypothetical model estimates central tendency within the observations well. The moderate value of d_r conveys this, while the negative values of E_1 and E understate the overall performance of the model.

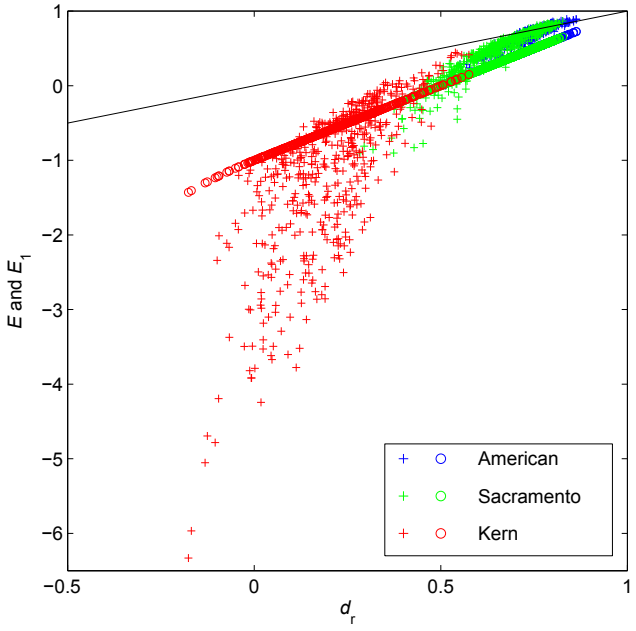

Fig. 3. Refined index of agreement (d_r) and the corresponding Nash–Sutcliffe coefficient of efficiency (E, colored "+" symbols) and Legates and McCabe value (E_1, colored "○" symbols) for each of the 500 different sets of model parameters, for each of three major California river basins; the American (+ and ○), Sacramento (+ and ○) and Kern (+ and ○). Data are the observed and modeled monthly streamflow from January 1950 to December 2005. The 1:1 line is plotted for reference. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

It also should be understood that d_r and E_1 are scaled algebraic descriptions of average error-magnitude, and that their main value lies in scientific interpretations of their magnitudes and signs. Their statistical (distributional) properties generally are of lesser importance. Nonetheless, if the original observed and corresponding model-predicted values are available, the pertinent distributional properties of d_r or E_1 can be assessed using resampling methods, such as bootstrapping (cf, Willmott et al., 1985).

3.1. Positive portions of E_1's and d_r's domains

Interpretations of either E_1 or d_r over the positive portion of its response domain are relatively straightforward (cf, Table 1); that is, each measure describes the relative extent (in a proportion) to which a set of model predictions is, on average, error free. A value of d_r of 0.75, for instance, indicates that 75 percent of the total or average reference error is accounted for by the model predictions.

3.2. Negative portions of E_1's and d_r's domains

Interpretations of values of E_1 or d_r that fall within the negative portion of its response domain are slightly less straightforward than interpretations of positive values, and often require some minor translation. Within the negative portion of d_r's domain, the relative magnitudes of each d_r as well as differences (distances) between two or more values of d_r can be somewhat more easily assessed and interpreted than negative values of E_1, primarily because d_r approaches its lower limit (−1.0) asymptotically.

Well-informed assessments of the predictive abilities of poorly performing models (MAE > c·MAD) require fidelity over the negative portion of the statistic's domain. Negative values of d_r (cf, Table 1) indicate that the average error-magnitude (MAE) is larger than the average reference error (2·MAD). The magnitude of each negative value of d_r, in addition, conveys the relative extent to which the average reference error underrepresents the actual average-error magnitude. When $d_r = -0.2$, for example, one interpretation is that the reference-error average underrepresents the average error-magnitude by 20 percent or, alternatively, is $(1 + d_r)$ times (80 percent of) the MAE. It also is true that the MAE is $[(1 + d_r)^{-1}]$ times larger than 2·MAD. Comparison between two negative values of d_r may be useful as well. When one model, for example, posts a d_r of −0.5 while another, less-accurate model obtains a d_r of −0.75, a fair interpretation is that the reference-error average (2·MAD) underrepresents the first MAE by 50 percent whereas, in the second case ($d_r = -0.75$), the underrepresentation is 75 percent. Such interpretations and comparisons are enhanced because d_r is constrained to asymptotically approach −1.0.

Legates and McCabe's (1999) E_1, on the other hand, can decrease substantially over the negative part of its response domain (cf, Table 1) and, as with d_r, interpretations of its negative values require some translation. Assessing and interpreting $[(|E_1| + 1)^{-1}]$, for instance, indicates the proportion (or percent) of the MAE that is comprised of the MAD, while $(|E_1| + 1)$ represents the number of multiples that the MAE is larger than the MAD. When $E_1 = -3.0$, for example, the MAD is 25 percent of the MAE while the MAE is 4 times larger than the MAD. Comparisons among differing negative values of E_1 can be made although interpretation may be somewhat less certain, because of E_1's unboundedness. Consider that, as the magnitudes of negative values of E_1 become increasingly large, they may become less stable and/or less meaningful, with seemingly large differences arising from only small or trivial changes within the average errors or deviations. Mathevet et al. (2006) observed a comparable issue with negative values of E, where a few highly negative values sufficiently skewed multi-basin distributions of efficiencies (E values) that meaningful statistics of the set of basin E values could not be obtained.

Competing values of d_r (or differences between competing values of d_r) can be usefully graphed and interpreted. Such graphical comparisons of differences between competing values of d_r are illustrated in two recent comparisons of models made by Ward et al. (2013) and Gaitan et al. (2014). Once again, owing to the behavior of E_1 over the unbounded negative portion of its domain (Willmott et al., 2012), comparing competing values of E_1 is more difficult.

3.3. Scaling of E, E_1 and d_r: the value of c

Willmott et al. (2012) proffered that the scaling coefficient (c) should be 2.0, rather than the 1.0 implied by Nash and Sutcliffe (1970) and Legates and McCabe (1999, 2013). The reasoning was and is based on the fact that each error $(P_i - O_i)$ is comprised conceptually of two deviations about central tendency, usually about the observed mean. A single error, in other words, can be characterized as the difference between two deviations about the observed mean, i.e. $[(P_i - \overline{O}) - (O_i - \overline{O})]$. Furthermore, even though the means cancel, 2n of these *deviations* yield n errors which, when summed, equal the total error. The denominators in Equation (2) and in part 1 of Equation (3) likewise should be comprised of n squared errors or n errors; in which case, it ($4 \cdot s_o^2$ or 2·MAD) becomes a conceptually balanced benchmark; and, importantly here, 2·MAD is what we refer to in this paper as the *reference-error average*. Of its elements, we refer to $(|O_i - \overline{O}|)$ as a *reference deviation* and to $(|O_i - \overline{O}| + |O_i - \overline{O}|)$ as a *reference error*.

Each reference error can be interpreted as a likely maximum distance that a perfect-model prediction, but with a sign error (i.e., $P_i = -O_i$), could be away from an actual observation. It is assumed that O_i is observed without error and that the perfect-model prediction would be equal to O_i. A reasonable estimate of the magnitude of a model-predicted deviation also is thought to be contained in the magnitude of an observed (reference) deviation. The sum of two corresponding reference deviations, in turn, is considered to represent a likely distance between the prediction of a very poor model (one that covaries negatively with observations) and observation. Although it may be argued—on an individual prediction-deviation basis—that our model-predicted-deviation (reference-deviation) assumption is imperfect, we believe that our estimated distribution of model-predicted deviations (reference deviations) is more reasonable. Taken together, these reference-deviation and -error estimates form an envelope or distribution of reference-error variability. Our reference errors, in turn, track twice the magnitudes of the observed deviations, and their position and scale parameters follow the scaled observed parameters. The sum of these n reference errors is the *reference-error total*

Table 1
Hypothetical sets (cases) of average error magnitude (MAE), average observed deviation magnitude (MAD) and corresponding values of E_1 and d_r. Our refined index of agreement (d_r) is evaluated separately over the positive (d_r^+) and negative (d_r^-) portions of its response domain.

MAE	MAD	$E_1 = 1 - \frac{\text{MAE}}{\text{MAD}}$	$d_r^+ = 1 - \frac{\text{MAE}}{2 \cdot \text{MAD}}$	$d_r^- = \frac{2 \cdot \text{MAD}}{\text{MAE}} - 1$
0.0	10.0	1.0	1.0	
5.0	10.0	0.5	0.75	
10.0	10.0	0.0	0.5	
15.0	10.0	−0.5	0.25	
20.0	10.0	−1.0	0.0	
25.0	10.0	−1.5		−0.2
40.0	10.0	−3.0		−0.5
80.0	10.0	−7.0		−0.75
200.0	10.0	−19.0		−0.9
400.0	10.0	−39.0		−0.95

and, conveniently, the *reference-error average* is 2·MAD. The logic is similar to that behind the "potential error" proposed by Willmott and Wicks (1980) and Willmott (1981). And, once again, the reference-error total and average are each comprised of $2n$ deviations. This reasoning leads to $c = 2.0$; not $c = 1.0$. We note that, even if one uses $c = 1.0$ with d_r, d_r is more readily interpretable than is either E or E_1 because of the finite bound on the negative portion of d_r's domain.

In contrast to d_r, the fractional parts of E and E_1 are scaled by n errors over $n/2$ reference errors or 2.0. This makes E and E_1 conceptually too low by a factor of 2.0. This also has the deleterious effect of increasing the chance that an E or E_1 value will fall within the unbounded negative portion of E or E_1's domain, discussed below.

4. Origin forcing and baselines for E, E_1 and d_r

Values of E and E_1 are zero, when the model predicts the observed mean ($\overline{O}$) as the estimate for each and every observation. Legates and McCabe (2013), for instance, believe that E_1 is "… preferable to many other statistics … [including d_r] … because [E_1 has] … a fundamental meaning at zero." Our interpretation differs, however, in part because their origin ($E_1 = 0.0$ baseline, when $P_i = \overline{O}$ for all i) can be too low, for reasons we describe below. We note that $d_r = 0.5$, when $P_i = \overline{O}$ for all i.

4.1. Two origin forcings: $P_i = \overline{O}$ or $P_i = \overline{O} \pm 2 \cdot MAD$, for all i

From an interpretational standpoint, $d_r = 0.5$, when $P_i = \overline{O}$ for all i, can be preferable to having $E_1 = 0.0$ with $P_i = \overline{O}$. Forcing E or E_1 or another index to zero, when $P_i = \overline{O}$ for all i, understates an important characteristic of such a model—that is, its ability to predict values in the exact center of a set of independent (of the model) observations (on average, no under- or over-prediction). We suspect that such an origin forcing ($E = 0.0$ or $E_1 = 0.0$, when $P_i = \overline{O}$, for all i) tacitly arises from the common use of covariance-based approaches to model building, wherein the mean of a fit function is forced to equal an observed mean. Our d_r is intended to be used to evaluate the predictions of a wider array of models than statistically fit models, especially numerical models of hydrologic and other environmental processes. As the observations to which the model estimates are to be compared should not inform these models, an index value of zero, when $P_i = \overline{O}$, for all i, underrepresents the ability of an unbiased model to estimate the independently observed variable. Reliable models should well estimate the first few empirical moments of the independently observed variable and especially its first moment.

Our assertion that d_r more faithfully communicates correspondence between observed and predicted first moments can be illustrated with a hypothetical example (Fig. 2). For purposes of this example, it is postulated that the domain of the variable of interest ranges from zero to 100, but that available observations only span a limited portion of this domain. Predictions corresponding to these observations by a hypothetical model are plotted against the observations in an observed-versus-predicted two-space. It is clear (Fig. 2) that this hypothetical model fairly well estimates central tendency within these observations and that the corresponding value of d_r ($d_r \approx 0.45$) adequately conveys this. The corresponding values of E and E_1 ($E \approx -0.46$ and $E_1 \approx -0.09$), on the other hand, understate the predictive ability of this model.

A model's ability to correctly estimate the first empirical moment of an independently observed variable is nontrivial skill which, in this case, $d_r \approx 0.5$ better represents intuitively. Not only does $d_r = 0.5$ better convey skill when $P_i = \overline{O}$, for all i, but when $d_r = 0.0$ each and every P_i differs from the independently observed mean by the reference-error average ($P_i = \overline{O} \pm 2 \cdot MAD$). Our origin forcing of one average reference error from the mean better indicates the model's inability to correctly reproduce observed central tendency.

4.2. Baseline adjustments: $f(\ )$ replaces $\overline{O}$

Seibert (2001), Legates and McCabe (2013) and many others make a point that the observed mean ($\overline{O}$) can be an inadequate baseline for measures like E_1 and d_r, and that other baselines, with more realistic variability, may be preferable. They say, for example, that "… the observed mean is not likely the most appropriate [baseline] choice …" and "… that climatologists … must strongly consider comparing their models against more appropriate baselines …" When a model estimates time-series values, for instance, it may be useful to replace the overall mean of the observed time series ($\overline{O}$) with observed seasonal or monthly values. Oyler et al. (2015) applied such baseline adjustments in order to reduce the number of high values of d_r that arose from their model successfully explaining well-known seasonal-cycle variability in daily air temperature.

While we agree that baseline adjustments or "benchmarks" (Seibert, 2001) in place of $\overline{O}$ ought to be evaluated when appropriate, we also think that baseline comparisons with $\overline{O}$ should *always* be made, interpreted and reported, even when other baseline adjustments are made. This is because index values computed with alternate benchmark functions, perhaps derived from different sets of observations, may not be sufficiently documented or reproducible and, in turn, meaningfully comparable; whereas $\overline{O}$ is well understood and comparisons of indices based on $\overline{O}$ can be made with relative confidence. Using more complicated baseline adjustments or benchmarks, in other words, may make interpretations and comparisons of indices more difficult, because each different benchmark is, in effect, an implicit rescaling of a model-performance index. In some cases, however, using a clear-cut benchmark—such as a persistence model—may produce readily interpretable results.

Such benchmark-dependent rescaling also is potentially more of an issue with E and E_1. As the benchmarks used are more representative of the observed variability than is the observed mean, the values of E and E_1 will be smaller and a greater number of them will fall within the negative and unbounded portions of E and E_1's response domains, making straightforward interpretations and meaningful comparisons of the performances of different models or different benchmarks more troublesome. Interpretational difficulties associated with E also may be exacerbated since the relative influence of each new squared deviation (from a new benchmark) on the sum of squared deviations may change in a counterintuitive way. Difficulty in interpreting the meaning of differences between the negative and other index values in Legates and McCabe's (1999) Table 1 illustrates these problems.

5. Comparative applications of E, E_1 and d_r to assess model streamflow predictions

The Soil and Water Assessment Tool (SWAT) hydrologic model (Arnold et al., 1998) applied to the Sierra Nevada range in California (Fig. 1) is used to further explore the similarities and differences among E, E_1 and d_r. SWAT is a basin-scale model designed to simulate the entire hydrologic cycle, including surface runoff, snowmelt, lateral soil flow, evapotranspiration, infiltration, deep percolation, and groundwater return flows, at the sub-basin scale. The Sierra Nevada SWAT model contains 379 sub-basins, wherein each sub-basin contributes runoff to a larger basin (14 basins in total for this model). Additional details about the SWAT model can

be found in Neitsch et al. (2005). Full details on this specific Sierra Nevada SWAT model, including input datasets and streamflow calibration data, can be found in Ficklin et al. (2012). Its subdomains include the modeled outlets of western slopes of the Sierra Nevada that span the rivers entering reservoirs from the Sacramento River to the north and the Kern River to the south (see Fig. 1 in Ficklin et al., 2012 for a more-detailed map of the study area and sub-basins). For our purposes, the model was run at a monthly time step from 1950 to 2005.

Similar to a sensitivity analysis—to determine which changes in model parameters result in large changes in streamflow—we allow model parameters (b_j values) to vary within a physically meaningful range (Table 2):

$$b_j\text{: } b_{j,abs\_min} \le b_j \le b_{j,abs\_max}$$

where b_{j,abs_min} is the model parameter minimum, b_{j,abs_max} is the model parameter maximum and j is the index of the model parameter set, varying from 1 to m, to produce m model parameter sets. Once the number of parameters for calibration as well as the minimum and maximum physically-meaningful ranges are established, Latin Hypercube sampling (McKay et al., 1979) using a uniform probability distribution for each parameter is carried out, leading to m different sets of model parameter combinations (with $m = 500$ for this work). It is important to note that the goal of this exercise is not model calibration, but to assess the sensitivities of the model-performance measures to multiple parameter sets. The Sierra Nevada SWAT model was then run for each of the m parameter combinations to allow m comparisons of observed and simulated streamflow for each of three basins. Based on prior knowledge of the region and previous SWAT model experience, we varied 26 parameters relating to surface and groundwater hydrology, as well as snowpack generation and snowmelt (Table 2).

Observed and modeled monthly streamflows from all 500 model runs were used to generate values of E, E_1 and d_r for each basin for the 1950–2005 period. Of the basins modeled (14 in total), we assess the properties of E, E_1 and d_r for the American, Sacramento and Kern River basins, based on these 500 model simulations. The American, Sacramento and Kern River basins were selected for analysis based on their differences in drainage area, elevation, slope and runoff coefficients (see Table 1 in Ficklin et al., 2012; for details). The Kern is the driest and most southerly basin, with low but variable flow. For the three basins, d_r is confined to the range of −0.176 to 0.863 while E and E_1 range from −6.33 to 0.895 and −1.43 to 0.726, respectively (Fig. 3). It is clear that d_r and E are closely related, but in a somewhat inconsistent way due to the squaring of both the error and observed variability terms in E. As expected, there is a monotonic and functional relationship between d_r and E_1 (see Fig. 3 in Willmott et al., 2012). In fact, when d_r is positive, $E_1 = 2d_r - 1$, so E_1 can never exceed d_r (and E_1 becomes negative when $d_r < 0.5$). When d_r is negative, the relationship with E_1 is nonlinear (Willmott et al., 2012), but very few (1.8%) of the model-parameter sets used here produced negative values of d_r; and only for the Kern Basin. Many of the model-parameter sets, however, produce negative values of E and E_1: 34.0% and 35.4% respectively. As the negative portions of the domains for E and E_1 are unbounded, interpretations of the magnitude of these values become problematic and this remains an important limitation for both of those statistics.

To further illustrate the properties and interpretations of d_r, we analyze the two model-parameter solutions that produced the lowest and highest values of d_r for the Kern River basin; what we will call the “worst” and “best” models. Scatterplots of observed and predicted monthly streamflow for the “worst” model show clear problems of model over-prediction when observed values are near-zero and model under-prediction for some of the highest monthly values in the Kern River basin (Fig. 4). All three model-evaluation statistics—d_r, E, and E_1—are negative for the worst model solution. Because observed streamflow within the Kern River basin is relatively low, but quite variable—and the variability term appears in the denominator of both E and E_1 (it does not for d_r when it is negative)—relatively small variations in the overall error

Table 2
Parameters of the basin-scale streamflow model (SWAT) that were randomly varied between their minimum (b_{j,abs_min}) and maximum (b_{j,abs_max}) values to obtain 500 different model-parameter solutions for sub-basins in the Sierra Nevada range.

Parameter	Minimum	Maximum	Description	
ALPHA_BF	0	1	Baseflow recession constant (unitless)	Groundwater
ALPHA_BNK	0	1	Baseflow alpha factor for bank storage (unitless)	
GW_DELAY	0	500	Groundwater delay time (days)	
GW_REVAP	0.02	0.2	Groundwater revap coefficient for movement of groundwater to unsaturated zone (unitless)	
GWQMIN	0	5000	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)	
LAT_TTIME	0	180	Lateral soil water travel time (days)	
REVAPMN	0	500	Threshold depth of water in the shallow aquifer for revap or percolation to the deep aquifer to occur (mm)	
CN	−15	15	Curve Number (% change from default value)	Surface water
SLSOIL	−15	15	Slope of soil (% change from default value)	
SOL_AWC	−15	15	Available water content of soil (% change from STATSGO value)	
SOL_K	−15	15	Saturated hydraulic conductivity of soil (% change from STATSGO value)	
SOL_BD	−15	15	Bulk density of soil (% change from STATSGO value)	
CH_N2	0	0.3	Manning's “n” value for the channel (unitless)	
CH_K2	0	150	Effective hydraulic conductivity of channel alluvium (mm/hr)	
EPCO	0	1	Plant uptake compensation factor (unitless)	
ESCO	0	1	Soil evaporation compensation factor (unitless)	
SURLAG	1	24	Surface runoff lag coefficient (days)	
PLAPS	0	1000	Precipitation lapse rate for elevation differences (mm H_2O/km)	Snowpack/snowmelt
TLAPS	−10	−0.1	Temperature lapse rate for elevation differences (°C/km)	
SFTMP	−5	5	Snowfall temperature (°C)	
SMFMN	0	10	Melt factor of snow on December 21st (mm H_2O/°C-day)	
SMFMX	0	10	Melt factor of snow on June 21st (mm H_2O/°C-day)	
SMTMP	−5	5	Snowmelt base temperature (°C)	
SNO50COV	0.01	0.9	Fraction of snow volume that represents 50% snow cover (unitless)	
SNOCOVMX	0	500	Minimum snow water content that corresponds to 100% snow cover (mm H_2O)	
TIMP	0.01	1	Snowpack temperature lag factor (unitless)	

*STATSGO: State Soil Geographic Data.

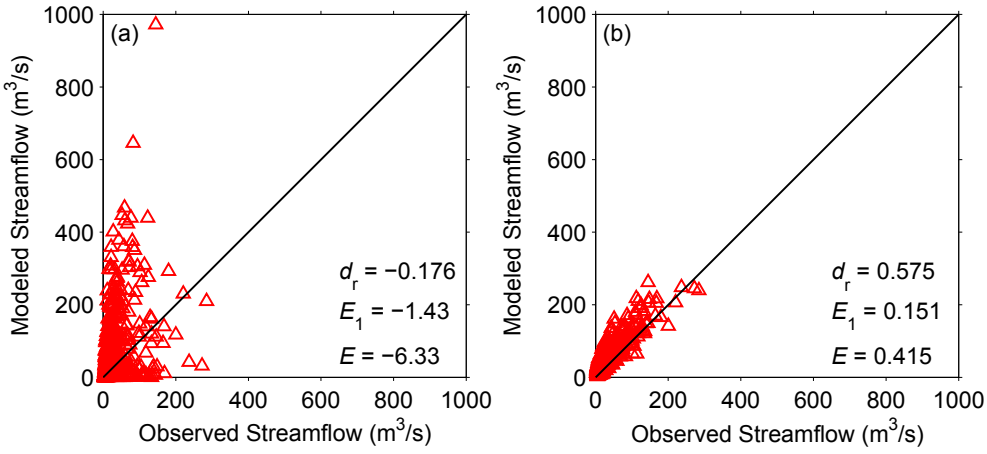

Fig. 4. Observed and modeled monthly streamflow for the Kern River basin from January 1950 to December 2005. The two sets of model parameters that produced the lowest (left) and highest (right) d_r values are shown. The 1:1 line and values of d_r, E_1, and E are provided for both cases.

of poor models can result in wide fluctuations in these unbounded statistics. The "best" model produces positive values for d_r, E, and E_1: 0.575, 0.415, and 0.151 respectively. Overall, the "best" model-parameter solution slightly overestimates some of the lower streamflow values while accurately simulating the higher values, so the relatively low value of E_1 is somewhat counter-intuitive. As outlined above, we argue that this is because d_r is more indicative of average model-performance errors than are values of E_1 or E (and E has the additional limitation associated with the squaring of error and observed variability terms).

6. Concluding remarks

Within this paper, we explain and analyze pertinent characteristics of three dimensionless model-performance statistics; Nash and Sutcliffe's (1970) E, Legates and McCabe's (1999) E_1 and Willmott et al.'s (2012) d_r. The behavior of E also has broader implications because it is typical of the class of measures sometimes referred to as benchmark measures (Seibert, 2001). Conceptual and algebraic comparisons are made among these three statistics and our assessments are supported by comparisons of how well each of these three indices are able to describe the predictive abilities of a set of re-parameterized versions of the SWAT streamflow model applied to the American, Sacramento and Kern River basins within California.

Even though the three model-performance statistics considered here are similar in form and dimensionless, we show that interpretations of d_r often are more straightforward than interpretations of E or E_1, giving d_r wider applicability. Its (d_r's) scaling has a credible conceptual basis, and its origin ($d_r = 0.0$, when $P_i = \overline{O} \pm 2 \cdot \mathrm{MAD}$, for all i) allows models that well-estimate central tendency to receive a higher score than does E or E_1 ($E = 0.0$ and $E_1 = 0.0$, when $P_i = \overline{O}$). We also observe that, if one has good reason to use an alternate (to $c = 2.0$) scaling of d_r or to use a more realistically varying benchmark than $\overline{O}$, the algebraic structure of d_r easily accommodates this. Lastly, the negative portions of E's and E_1's domains are unbounded, while all of d_r's variability occurs within a bounded (by 1.0 and −1.0) domain. This makes visualizations and interpretations of negative values of d_r— associated with poorly performing models—more straightforward.

Our findings indicate that E_1 and d_r are superior to E and its related benchmark measures (Seibert, 2001); that is, when used in assessing model performance rather than in model-training efforts. The reason is that average-error and agreement measures which are based on sums of error magnitudes are, in general, superior to comparable measures based on sums of squared errors (Willmott and Matsuura, 2005, 2006; Willmott et al., 2009). Our examinations of and comparisons among d_r, E_1 and E underscore the importance of evaluating and reporting a meaningful dimensionless measure of model performance in virtually all model-assessment or -comparison studies. Of the three measures considered, our findings indicate that d_r is more versatile and readily interpretable. It (d_r) can be applied to virtually any set of pairwise model-predicted and observed values (having comparable units) and values of d_r can be usefully compared across studies.

Acknowledgements

Aspects of the research reported on in this paper were made possible by NASA grant NNX12AJ20G (Sub Award number 12-001JNA to the University of Delaware through Delaware State University) and we are most grateful for this support. The authors also gratefully acknowledge financial support for this work from the U.S. Environmental Protection Agency through EPA STAR Grant No. RD-83419101-0. Three reviews of our manuscript were very helpful and we greatly appreciate the effort these reviewers made to improve our paper.

References

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment Part I: model development. J. Am. Water Resour. Assoc. 34, 73–89.

Ashofteh, P.-S., Haddad, O.B., Mariño, M.A., 2015. Risk analysis of water demand for agricultural crops under climate change. J. Hydrol. Eng. 20 (4), 04014060-1–04014060-10.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Berthet, L., Andreassian, V., Perrin, C., Loumagne, C., 2010. How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. Hydrol. Sci. J. 55 (6), 1063–1073.

Bring, A., Destouni, G., 2014. Arctic climate and water change: model and observation relevance for assessment and adaptation. Surv. Geophys. 35, 853–877. http://dx.doi.org/10.1007/s10712-013-9267-6.

Crochemore, L., Perrin, C., Andreassian, V., Ehret, U., Seibert, S., Grimaldi, S., Gupta, H., Paturel, J., 2015. Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrol. Sci. J. 60 (3), 402–423.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third ed. Wiley, New York.

Ehret, U., Zehe, E., 2011. Series distance—an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events.

Hydrol. Earth Syst. Sci. 15, 877–896.
Ficklin, D.L., Stewart, I.T., Maurer, E.P., 2012. Projections of 21st century Sierra Nevada local hydrologic flow components using an ensemble of general circulation models. J. Am. Water Resour. Assoc. 48, 1104–1125.
Gaitan, C.F., Hsieh, W.W., Cannon, A.J., 2014. Comparison of statistically downscaled precipitation in terms of future climate indices and daily variability for southern Ontario and Quebec, Canada. Clim. Dyn. 43, 3201–3217. http://dx.doi.org/10.1007/s00382-014-2098-4.
Gupta, H.V., Kling, H., Koray, K., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91.
Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241.
Legates, D.R., McCabe, G.J., 2013. A refined index of model performance: a rejoinder. Int. J. Climatol. 33 (4), 1053–1056.
Mathevet, T., Michel, C., Andréassian, V., Perrin, C., 2006. A bounded version of the Nash–Sutcliffe criterion for better model assessment on large sets of basins. In: Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment–MOPEX, pp. 211–219. IAHS Publ. 307.
McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, I. A discussion of principles. J. Hydrol. 10, 282–290.
Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., King, K.W., 2005. Soil and Water Assessment Tool Theoretical Documentation: Version 2005. Texas Water Resources Institute, College Station, Texas.
Oyler, J.W., Ballantyne, A., Jencso, K., Sweet, M., Running, S.W., 2015. Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. Int. J. Cimatol. 35, 2258–2279. http://dx.doi.org/10.1002/joc.4127.
Seibert, J., 2001. On the need for benchmarks in hydrological modelling. Hydrol. Process. 15, 1063–1064.
Singh, V.P., Frevert, D.K. (Eds.), 2006. Watershed Models. CRC Press, Boca Raton, FL.
Ward, E.J., Bell, D.M., Clark, J.S., Ram, O., 2013. Hydraulic time constants for transpiration of loblolly pine at a free-air carbon dioxide enrichment site. Tree Physiol. 33, 123–134. http://dx.doi.org/10.1093/treephys/tps114.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194.
Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63, 1309–1313.
Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90 (c5), 8995–9005.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82.
Willmott, C.J., Matsuura, K., 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. Int. J. Geogr. Inf. Sci. 20, 89–102.
Willmott, C.J., Matsuura, K., Robeson, S.M., 2009. Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ. 43, 749–752.
Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. Int. J. Climatol. 32, 2088–2094.
Willmott, C.J., Wicks, D.E., 1980. An empirical method for the spatial interpolation of monthly precipitation within California. Phys. Geogr. 1, 59–73.