INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **32**: 2088–2094 (2012)
Published online 9 September 2011 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.2419

RMetS
Royal Meteorological Society

Short Communication
A refined index of model performance

Cort J. Willmott,[a]* Scott M. Robeson[b] and Kenji Matsuura[a]
[a] *Center for Climatic Research, Department of Geography, University of Delaware, Newark, DE 19716, USA*
[b] *Department of Geography, Indiana University, Bloomington, IN 47405, USA*

ABSTRACT: In this paper, we develop, present and evaluate a refined, statistical index of model performance. This 'new' measure (d_r) is a reformulation of Willmott's index of agreement, which was developed in the 1980s. It (d_r) is dimensionless, bounded by −1.0 and 1.0 and, in general, more rationally related to model accuracy than are other existing indices. It also is quite flexible, making it applicable to a wide range of model-performance problems. The two main published versions of Willmott's index as well as four other comparable dimensionless indices – proposed by Nash and Sutcliffe in 1970, Watterson in 1996, Legates and McCabe in 1999 and Mielke and Berry in 2001 – are compared with the new index. Of the six, Legates and McCabe's measure is most similar to d_r. Repeated calculations of all six indices, from intensive random resamplings of predicted and observed spaces, are used to show the covariation and differences between the various indices, as well as their relative efficacies.

KEY WORDS accuracy indices; model-performance statistics

Received 4 February 2011; Revised 22 July 2011; Accepted 23 July 2011

1. Introduction

Numerical models of climatic, hydrologic, and environmental systems have grown in number, variety and sophistication over the last few decades. There has been a concomitant and deepening interest in comparing and evaluating the models, particularly to determine which models are more accurate (e.g. Krause *et al*., 2005). Our interest lies in this arena; that is, in statistical approaches that can be used to compare model-produced estimates with reliable values, usually observations.

Our main purpose in this paper is to present and evaluate a refined version of Willmott's dimensionless index of agreement (Willmott and Wicks, 1980; Willmott, 1981, 1982, 1984; Willmott *et al*., 1985). The refined index, we believe, is a nontrivial improvement over earlier versions of the index and is quite flexible, making it applicable to an extremely wide range of model-performance applications. Our discussion contains a brief history, a description and assessment of its form and properties, and comparisons with a set of other dimensionless measures of average model accuracy to illustrate its relative effectiveness.

2. Background

Statistical measures of model performance, including the index of agreement, commonly compare model estimates or predictions (P_i; $i = 1, 2, \ldots, n$) with pairwise-matched observations (O_i; $i = 1, 2, \ldots, n$) that are judged to be reliable. The units of P and O should be the same. The set of model-prediction errors usually is composed of the $(P_i - O_i)$ values, with most dimensioned measures of model performance being based on the central tendency of this set.

The majority of dimensionless measures of average error, once again including the index of agreement, are framed as

$$\rho = 1 - \delta/\mu \tag{1}$$

where δ is a dimensioned measure of average error or, more precisely, average error-magnitude and μ is a basis of comparison. The selection of δ, of course, determines how well the average error-magnitude will be represented, while the choice of μ determines the lower limit of Equation (1), as well as the sensitivity of ρ to changes in δ. As $\delta \geq 0$ and $\mu > 0$, the upper limit of Equation (1) is 1.0 and indicates perfect model performance. In most cases, μ is defined such that the lower limit of Equation (1) is 0, −1, or $-\infty$.

3. Brief history of the index of agreement

The original form of Willmott's index of agreement (Willmott and Wicks, 1980; Willmott, 1981) was a specification of Equation (1). Willmott and Wicks used d to represent the index (rather than ρ) and we will follow their convention here. It (d) was a sums-of-squares-based

* Correspondence to: Cort J. Willmott, Center for Climatic Research, Department of Geography, University of Delaware, Newark, DE 19716, USA. E-mail: willmott@udel.edu

measure, within which δ was the sum of the squared errors while μ was the overall sum of the squares of sums of the absolute values of two partial differences from the observed mean, $|P_i - \overline{O}|$ and $|O_i - \overline{O}|$. Thus, the form of the original index was

$$d = 1 - \frac{\sum_{i=1}^{n}[(P_i - \overline{O}) - (O_i - \overline{O})]^2}{\sum_{i=1}^{n}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \quad (2a)$$

which simplifies to and is commonly written as

$$d = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \quad (2b)$$

This specification of μ ensured that the potential upper limit of δ was μ and, in turn, that the lower limit of d (indicating 'complete disagreement') was zero. Willmott reasoned that d should describe the relative covariability of P and O about an estimate of the 'true' mean; that is, about $\overline{O}$. This not only 'conveniently' bounds Willmott's d on the lower end, but when $d = 0$, it can be physically meaningful. When all the P_i s equal $\overline{O}$, for example, $d = 0$. Moreover, when every P_i and O_i pair is on the opposite side of $\overline{O}$, $d = 0$; in other words, the model-predicted variability about $\overline{O}$ behaves inversely to the observed variability ($O_i - \overline{O}$). Model predictions that are completely out of phase with observations might produce such a response.

As an increasing number of applications of d to model-performance problems were made, Willmott and his graduate students began to suspect that squaring the errors, prior to summing them, was a problem (Willmott, 1982, 1984; Willmott *et al*., 1985). It was recognized that the larger errors, when squared, over-weighted the influence of those errors on the sum-of-squared errors. The full nature and extent of this problem was not appreciated until the upper limit of the sum-of-squared errors was investigated more thoroughly (Willmott and Matsuura, 2005, 2006). Nonetheless, Willmott *et al*. (1985) put forward a version of d that was based upon sums of the absolute values of the errors (calling it d_1) and he and his collaborators have preferentially used d_1 ever since. The form of d_1 is

$$d_1 = 1 - \frac{\sum_{i=1}^{n}|P_i - O_i|}{\sum_{i=1}^{n}(|P_i - \overline{O}| + |O_i - \overline{O}|)} \quad (3)$$

An advantage that d_1 has over d is that it approaches 1.0 more slowly as the P approaches O and, therefore, provides greater separation when comparing models that perform relatively well. It (d_1) also is much less sensitive to the shape of the error-frequency distribution and, as a consequence, to errors concentrated in outliers.

Our experience has been that average-error or- deviation measures that are based on absolute values of differences – like the mean-absolute error (MAE) and mean-absolute deviation (MAD) – are, in general, preferable to those based on squared differences, like the root-mean-squared error (RMSE) (Willmott and Matsuura 2005, 2006; Willmott *et al*., 2009), where $\mathrm{MAE} = n^{-1}\sum_{i=1}^{n}|P_i - O_i|$, $\mathrm{MAD} = n^{-1}\sum_{i=1}^{n}|O_i - \overline{O}|$, and $\mathrm{RMSE} = [n^{-1}\sum_{i=1}^{n}(P_i - O_i)^2]^{0.5}$. Dimensionless indices of model performance, in turn, also should be based on absolute values of differences or deviations. There are instances, however, when useful information can be gleaned from a set of squared differences. The RMSE, for instance, can be decomposed into systematic and unsystematic components (Willmott, 1981), and, when comparing mapped variables, the RMSE can help to distinguish between differences due to quantity and those due to location (Pontius *et al*., 2008).

4. A refined index of agreement

Our primary goal is to present a refined index of agreement that, like d and d_1, is bounded on both the upper and lower ends. Many other existing indices are bounded on the upper end (usually by 1.0) but lack a finite lower bound (cf. Legates and McCabe, 1999; Krause *et al*., 2005), which makes assessments and comparisons of poorly performing models difficult. With this in mind, we have developed our new index with an easily interpretable lower limit of -1.0. With an upper limit of 1.0, the range of our new index is double the range of d or d_1. Our refined index also is logically related to increases and decreases in MAE.

In retrospect, while the over-sensitivity of d to large error-magnitudes was reduced in d_1, two aspects of d_1 remain suboptimal. Its overall range ($0 \le d_1 \le 1$) remained somewhat narrow to resolve adequately the great variety of ways that P can differ from O. It also seems that including the variability of P (around $\overline{O}$) within μ makes the interpretation of μ somewhat murky. That is, as long as μ contains variability in P, it cannot be interpreted as a model-independent standard of comparison for δ.

A natural way to remove the influence of P on μ is to replace P with O within the denominator of d_1. The revised d_1 (d_1') then can be written as

$$d_1' = 1 - \frac{\sum_{i=1}^{n}|(P_i - \overline{O}) - (O_i - \overline{O})|}{\sum_{i=1}^{n}(|O_i - \overline{O}| + |O_i - \overline{O}|)}$$

$$= 1 - \frac{\sum_{i=1}^{n} |P_i - O_i|}{2\sum_{i=1}^{n} |O_i - \overline{O}|} \quad (4)$$

This revision, however, makes d_1' unbounded on the lower end, which undermines interpretations of index values associated with poorly performing models. Our solution is to invert the fractional part of the expression and subtract 1.0 from it, when d_1' falls below zero. This follows an approach taken by Willmott and Feddema (1992) to refine a climatic moisture index. Our new or refined index of agreement (d_r) then can be expressed as

$$d_r = \begin{cases} 1 - \dfrac{\sum_{i=1}^{n} |P_i - O_i|}{c\sum_{i=1}^{n} |O_i - \overline{O}|}, \text{ when} \\ \sum_{i=1}^{n} |P_i - O_i| \le c\sum_{i=1}^{n} |O_i - \overline{O}| \\ \dfrac{c\sum_{i=1}^{n} |O_i - \overline{O}|}{\sum_{i=1}^{n} |P_i - O_i|} - 1, \text{ when} \\ \sum_{i=1}^{n} |P_i - O_i| > c\sum_{i=1}^{n} |O_i - \overline{O}| \end{cases} \quad (5)$$

with $c = 2$, following Equation (4).

Interpretation of d_r is relatively straightforward. It indicates the sum of the magnitudes of the differences between the model-predicted and observed deviations about the observed mean relative to the sum of the magnitudes of the perfect-model ($P_i = O_i$, for all i) and observed deviations about the observed mean. A value of d_r of 0.5, for example, indicates that the sum of the error-magnitudes is one half of the sum of the perfect-model-deviation and observed-deviation magnitudes. When $d_r = 0.0$, it signifies that the sum of the magnitudes of the errors and the sum of the perfect-model-deviation and observed-deviation magnitudes are equivalent. When $d_r = -0.5$, it indicates that the sum of the error-magnitudes is twice the sum of the perfect-model-deviation and observed-deviation magnitudes. Values of d_r near -1.0 can mean that the model-estimated deviations about $\overline{O}$ are poor estimates of the observed deviations; but, they also can mean that there simply is little observed variability. As the lower limit of d_r is approached, interpretations should be made cautiously.

It also is possible to interpret d_r in terms of the average-error and -deviation measures, MAE and MAD. Our statistic (d_r) inversely follows a scaling ($1/c$) of MAE/MAD. With $c = 2$, the influence of the MAD is doubled; one of these two MADs accounts for the observed mean-absolute deviation and the other represents the average magnitude of the perfect-model deviations. Recall that each error-magnitude within the MAE represents the difference between two deviations, one model-predicted and one observed; thus, the number of deviations evaluated within the numerator and within the denominator of the fractional part of d_r are the same and in conceptual balance.

5. Comparison with other dimensionless measures of average error

Within this section of the paper, we assess d_r's responses to varying patterns of differences between P and O and compare them with the corresponding responses of several comparable measures. Comparable measures that we consider are: d, d_1, Watterson's M, Mielke and Berry's $\Re$, Nash and Sutcliffe's E and Legates and McCabe's E_1. It is common for the MSE (RMSE^2) or a closely related measure to represent δ (cf. Watterson, 1996) and, herein, d, M and E represent this class of measures. With increasing frequency, however, dimensionless indices based on the MAE have appeared in the literature (cf. Mielke, 1985; Willmott *et al.*, 1985; Mielke and Berry, 2001; Legates and McCabe, 1999; Krause *et al.*, 2005). This class of measures is represented here by d_1, $\Re$, E_1 as well as by d_r.

To show how each of the indices varies, relative to d_r, simulated values of P and O were created using a uniform random number generator. More specifically, random samples of size $n = 10$ were generated separately for P and for O (using a pseudo-random number generator). Since our primary interest is in identifying the envelopes of covariability among the various indices, a small sample size was preferred. Values of each measure or index of interest were calculated from the ten pairs of sampled values. To estimate the full extents of the envelopes of covariation between d_r and each of the other six measures, this random sampling and calculation process was repeated 100 000 times for each index. A stratified percentile-based subsample of these values is plotted to depict each envelope of covariability (Figures 1–3). Less intensive samples, with the mean value of P offset from that of O, are used to demonstrate the behaviours of the indices for the case of overprediction (Figure 4).

On each of our first three scatterplots, d_r is the x-axis variable and two of the other six indices are plotted along the y-axis (Figures 1–3). On the first graph (Figure 1), d and d_1 are plotted against d_r; on the second graph (Figure 2), M and $\Re$ are plotted against d_r; and, on the third graph (Figure 3), E and E_1 are plotted against d_r. So, within each scatterplot, a sum-of-squares-based and a sum-of-absolute-values-based measure are plotted against d_r.

It is clear that d_r differs substantially from its two predecessors, d and d_1 (Figure 1). Recall that d is a sum-of-squares-based measure and d_1 a sum-of-absolute-values-based measure. First and foremost, the range of d_r is twice that of d and d_1. For models with large error distributions (relative to variability in O and P), values of d or d_1 usually are higher than comparable

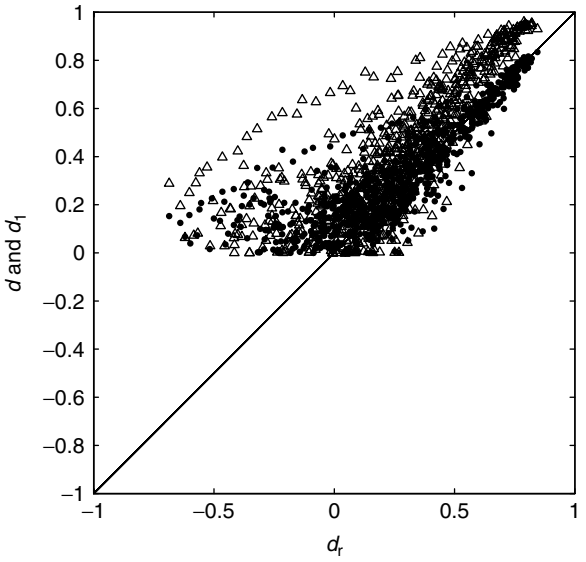

Figure 1. Stratified subsamples of statistics from the 100 000 pair-wise values of d_r and d (triangles) and of d_r and d_1 (black dots) are plotted. Each comparable value of each index was computed from the same $n = 10$ sample of P and O. The 1 : 1 line is plotted for reference.

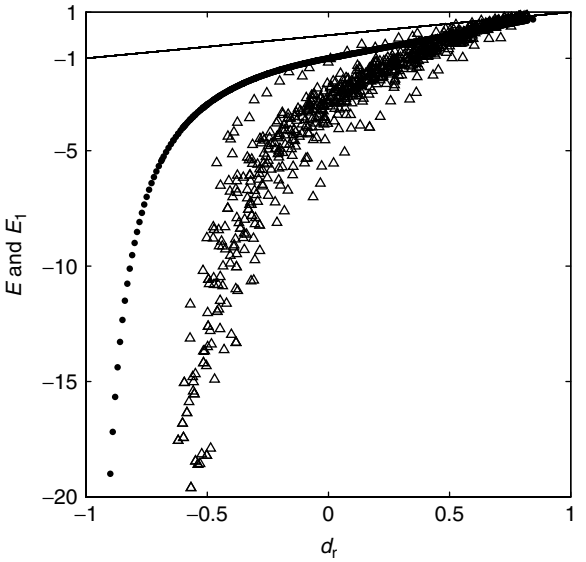

Figure 3. Stratified subsamples of statistics from the 100 000 pair-wise values of d_r and E (triangles) and of d_r and E_1 (black dots) are plotted. Each comparable value of each index was computed from the same $n = 10$ sample of P and O. The 1 : 1 line is plotted for reference.

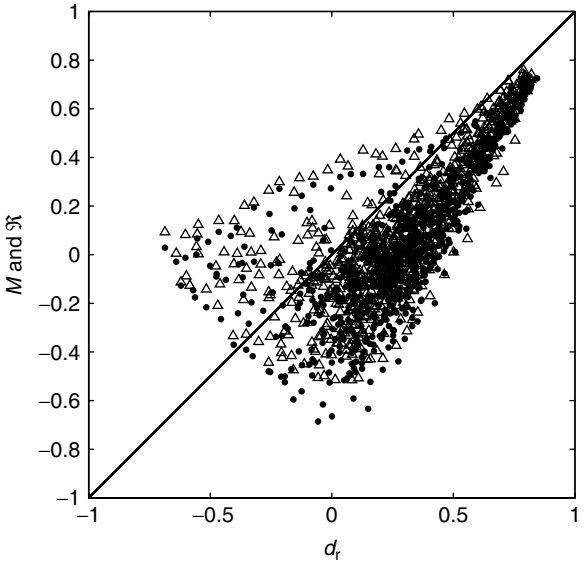

Figure 2. Stratified subsamples of statistics from the 100 000 pair-wise values of d_r and M (triangles) and of d_r and $\Re$ (black dots) are plotted. Each comparable value of each index was computed from the same $n = 10$ sample of P and O. The 1 : 1 line is plotted for reference.

values of d_r. Especially for d_1, this difference declines as P approaches O, to the point where d_r approaches d_1 for models with very small error distributions. There is more scatter (and high values) exhibited among the random sample-based values of d, relative to d_1, owing to the effects of squaring, especially on extrema and within the denominator of the fractional part. It also is apparent that increases and decreases in d_r are not monotonically related to increases and decreases in d and d_1. This is because P is a variable within the denominator of fractional parts of d and d_1, but not within d_r. Thus, d_r tends to behave rather differently – and more rationally – than either d or d_1.

Our comparisons among d_r, Watterson's (1996) M, and Mielke and Berry's (2001) $\Re$ (Figure 2) also show considerable differences among the three measures. The form of Watterson's index that we examine here is

$$M = (2/\pi)\sin^{-1}\left\{1 - \frac{\text{MSE}}{s_P^2 + s_O^2 + (\overline{P} - \overline{O})^2}\right\} \qquad (6)$$

where s_P snd s_O are the standard deviations of P and O, respectively. Mielke and Berry discuss several forms of their index; but, we only consider

$$\Re = 1 - \frac{\text{MAE}}{n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}|P_j - O_i|} \qquad (7)$$

Both scatterplots (M *versus* d_r and $\Re$ *versus* d_r) are centered not too far from zero, since each measure has a similar structure and their domains extend into negative numbers. Both indices exhibit a similar scatter pattern (this is not entirely expected, as M is a sum-of-squares-based measure and $\Re$ is not). As with d and d_1, increases and decreases in d_r are not monotonically related to increases and decreases in M and $\Re$ for the same reason: P is contained within the denominator of fractional parts of M and $\Re$.

Of the six comparable measures, the indices of Nash and Sutcliffe (1970) and Legates and McCabe (1999) are most closely related to d_r, because – within these measures – P does not appear in the denominators of

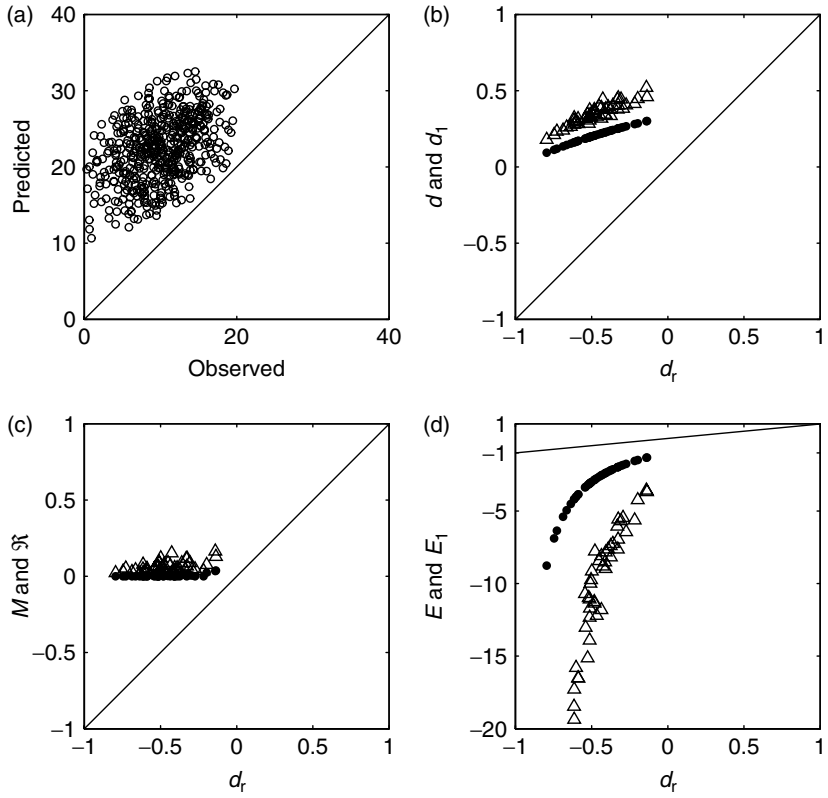

Figure 4. Demonstration of index values for the case of overprediction. Using uniform distributions, 500 values of O and P were generated (with O centered on 10 and P centered on 20). Fifty subsamples of size $n = 10$ are drawn, and pair-wise values of d_r and the other indices are calculated. Panels show: (a) 500 values of O and P, (b) 50 pair-wise values of d_r and d (triangles) and of d_r and d_1 (black dots), (c) 50 pair-wise values of d_r and M (triangles) and of d_r and $\Re$ (black dots), and (d) 50 pair-wise values of d_r and E (triangles) and of d_r and E_1 (black dots). In all cases, the 1 : 1 line is plotted for reference.

fractional parts. Nash and Sutcliffe's coefficient of efficiency (E) is

$$E = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \tag{8}$$

whereas Legates and McCabe's index (E_1) is written as

$$E_1 = 1 - \frac{\sum_{i=1}^{n}|P_i - O_i|}{\sum_{i=1}^{n}|O_i - \overline{O}|} \tag{9}$$

It is apparent in Equations (8) and (9) that E is based on the squares of differences while E_1 is based on the absolute values of differences. It also is clear that these two measures are similar to d_r, especially over the positive portion of its domain (Figure 3; note that Figure 3 has a different y-axis scaling from Figures 1 and 2). The coefficient of efficiency (E) is positively correlated with d_r and increasingly so as both measures approach their upper limits; but, the summing of squared differences within E precludes a monotonic relationship between the increases and decreases in d_r and in E. Legates and McCabe's measure, on the other hand, is monotonically and functionally related to our new index; and, when positive, E_1 is equivalent to d_r with $c = 1$. As mentioned above, we think that c = 2 is a better scaling, because it balances the number of deviations evaluated within the numerator and within the denominator of the fractional part. It (E_1) is an underestimate of d_r, as is evident in the functional relationship(s) between d_r and E_1. Over the positive portion of d_r's domain, $d_r = 0.5(E_1 + 1)$ while, when d_r is negative, $d_r = -[2(E_1 - 1)^{-1} + 1]$. The second expression also shows d_r's linearisation of E_1's exponential decline from 0 to $-\infty$.

A nontrivial difference between d_r and E_1, as well as between d_r and E, is the indices' behaviour over the negative portions of their domains. The magnitudes of both E_1 and E increase exponentially in the negative direction (Figure 3), which can make comparisons among some model estimates difficult. When the deviations around $\overline{O}$ are quite small or perhaps trivial, for instance, even small differences among competing sets of model estimates can produce substantially different values of E_1 or of E. In comparing models that estimate daily or monthly precipitation in an arid location, for example, relatively small differences between

the sets of model estimates could produce vastly different values of E_1 or of E. Values of d_r, on the other hand, would be more usefully comparable to one another.

It is clear that Legates and McCabe (and Nash and Sutcliffe before them) appreciated the importance of specifying μ with variation within the observed variable only. Legates and McCabe further understood the importance of evaluating error- and deviation-magnitudes, rather than their squares. Their measure (E_1), in turn, has a structure similar to that of d_r but with a substantially different scaling and lower limit, as discussed above.

To show the behaviour of the indices for a specific case of predicted *versus* observed data, we selected a typical pattern: overprediction (Figure 4). For this case, we show the scatterplot that we sample from (Figure 4(a)), as well as scatterplots of the other six measures *versus* d_r for 50 random samples of the pair-wise values of P and O. On each of the three scatterplots, d_r is the x-axis variable and two of the other six indices are plotted along the y-axis (i.e. Figure 4(b)–(d) have the same setup as Figures 1–3). It is clear that both d and d_1 are much less responsive than d_r to the various configurations of overprediction that can occur (Figure 4(b)). For this particular case, where the magnitude of MAE is consistently larger than the magnitude of the observed variability, d_r produces negative values while the values of d can range from 0.2 to over 0.5 (d_1 is more conservative than d but also is less responsive than d_r to the types of O *versus* P samples that are produced). For the 50 samples from our overprediction distribution, both M and $\Re$ produce almost no variation. $\Re$, in particular, is very close to zero for almost all of the varied samples within the overprediction example. Similar to Figure 3, Figure 4(d) demonstrates how small differences among the various observed and predicted samples can produce substantially different values of E_1 or of E that are difficult to interpret. It is useful to note that swapping O and P in this example (i.e. producing a case where the model systematically underpredicts) produces virtually no change in any of the indices. In cases where O and P have different magnitudes of variability, this symmetry of overprediction and underprediction does not occur.

6. Bases of comparison, other than $\overline{O}$

It is usual for indices of agreement to compare predicted and observed variability about $\overline{O}$, as $\overline{O}$ is often the best available estimate of the 'true' average. Sampling and other representational problems, however, may render $\overline{O}$ a suboptimal representation of central tendency in some circumstances. A better estimate of the true mean, for instance, may be one that varies over space and time, rather than one that is averaged over the entire domain. Consider that, when examining an observed time series of a climate variable, it may be better (more representative) to use observed seasonal means rather than the mean of the entire time series. Our refined index (d_r) can accommodate the replacement of $\overline{O}$ by any appropriate function (e.g. sub-period averages) of the observed variable.

7. Concluding remarks

A refined version of Willmott's dimensionless index of agreement was developed, described and compared with two previously published versions of Willmott's index of agreement. Our presentation also contained a brief history of the main forms of Willmott's index of agreement. In addition, the relative performance of the new index was compared with the performances of comparable indices proposed by Nash and Sutcliffe (1970), Watterson (1996), Legates and McCabe (1999) and Mielke and Berry (2001). The refined index appears to be a non-trivial improvement over earlier versions of the index as well as over other comparable indices. It is flexible, relatively well behaved, and applicable to a wide range of model-performance applications.

Variations in d_r's responses to patterns of differences between comparable sets of model-predicted values (P) and reliable observations (O) were examined. They also were compared with the corresponding responses of six comparable measures. Comparable measures that we considered were Willmott's d and d_1, Watterson's M, Mielke and Berry's $\Re$, Nash and Sutcliffe's E and Legates and McCabe's E_1. To examine how each of these indices varied, relative to d_r, simulated values of P and O were created using a uniform random number generator. Values of the indices then were calculated from small samples taken from the simulated values of P and O. This random sampling and calculation process was repeated 100 000 times for each index, and a stratified pair-wise subsample for each set of two indices was plotted to depict each bivariate envelope of covariability.

Our new measure shows no consistent relationship (monotonic increase or decrease) with five of the six other measures to which we compared it. It does share a functional relationship, however, with Legates and McCabe's (E_1) measure. Their measure (E_1) is monotonically related to our new index; and, when positive, E_1 would be equivalent to d_r if d_r were rescaled with $c = 1$. We argue in the paper that $c = 2$ is a preferred scaling. It also is true that E_1 always is an underestimate of d_r. Over the positive portion of d_r's domain, the underestimation is linear but, when d_r is negative, the magnitude of E_1's underestimation of d_r increases exponentially. Comparisons among some model estimates using E_1, in turn, can be problematic. When the deviations around the observed mean are quite small, for instance, even small differences among competing sets of model estimates can produce substantially different values of E_1. Values of d_r derived from competing sets of models should be usefully comparable to one another. It is important to point out; however, that both Nash and Sutcliffe and Legates

and McCabe preceded us in identifying the importance of including only observed deviation within the basis of comparison (their denominators) of the fractional part.

Acknowledgements

Several of the ideas presented in this paper are extensions of concepts previously considered by the authors of Willmott *et al.*, 1985. In particular, we are indebted to David Legates for his early recognition of the potential utility of an absolute-value-based version of Willmott's index of agreement. Aspects of the research reported on here were made possible by NASA Grant NNG06GB54G to the Institute of Global Environment and Society (IGES) and we are most grateful for this support.

References

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

Legates DR, McCabe GJ Jr. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**(1): 233–241.

Mielke PW Jr. 1985. Geometric concerns pertaining to applications of statistical tests in the atmospheric sciences. *Journal of Atmospheric Science* **42**: 1209–1212.

Mielke PW Jr, Berry KJ. 2001. *Permutation Methods: A Distance Function Approach*. Springer-Verlag: New York; 352.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology* **10**(3): 282–290.

Pontius Jr, RG, Thontteh O, Chen H. 2008. Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics* **15**(2): 111–142.

Watterson IG. 1996. Non-dimensional measures of climate model performance. *International Journal of Climatology* **16**: 379–391.

Willmott CJ. 1981. On the validation of models. *Physical Geography* **2**: 184–194.

Willmott CJ. 1982. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* **63**: 1309–1313.

Willmott CJ. 1984. On the evaluation of model performance in physical geography. In *Spatial Statistics and Models,* Gaile GL, Willmott CJ (eds). D. Reidel: Boston; 443–460.

Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. 1985. Statistics for the evaluation of model performance. *Journal of Geophysical Research* **90**(C5): 8995–9005.

Willmott CJ, Feddema JJ. 1992. A more rational climatic moisture index. *Professional Geographer* **44**(1): 84–88.

Willmott CJ, Matsuura K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **30**: 79–82.

Willmott CJ, Matsuura K. 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science* **20**(1): 89–102.

Willmott CJ, Matsuura K, Robeson SM. 2009. Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment* **43**(3): 749–752.

Willmott CJ, Wicks DE. 1980. An empirical method for the spatial interpolation of monthly precipitation within California. *Physical Geography* **1**: 59–73.

