A refined index of model performance: A rejoinder

Article *in* International Journal of Climatology · March 2013

DOI: 10.1002/joc.3487

David Legates
University of Delaware
116 PUBLICATIONS **14,697** CITATIONS

SEE PROFILE

Gregory J. Mccabe
United States Geological Survey
155 PUBLICATIONS **16,355** CITATIONS

SEE PROFILE

INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **33**: 1053–1056 (2013)
Published online 18 April 2012 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.3487

RMetS
Royal Meteorological Society

Short Communication

A refined index of model performance: a rejoinder

David R. Legates[a]* and Gregory J. McCabe[b]

[a] *Department of Geography, Center for Climatic Research, University of Delaware, Newark, DE 19716-2541, USA*
[b] *United States Geological Survey, Denver, CO 80225, USA*

ABSTRACT: Willmott *et al.* [Willmott CJ, Robeson SM, Matsuura K. 2012. A refined index of model performance. *International Journal of Climatology*, forthcoming. DOI:10.1002/joc.2419.] recently suggest a refined index of model performance (d_r) that they purport to be superior to other methods. Their refined index ranges from -1.0 to 1.0 to resemble a correlation coefficient, but it is merely a linear rescaling of our modified coefficient of efficiency (E_1) over the positive portion of the domain of d_r. We disagree with Willmott *et al.* (2012) that d_r provides a better interpretation; rather, E_1 is more easily interpreted such that a value of $E_1 = 1.0$ indicates a perfect model (no errors) while $E_1 = 0.0$ indicates a model that is no better than the baseline comparison (usually the observed mean). Negative values of E_1 (and, for that matter, $d_r < 0.5$) indicate a substantially flawed model as they simply describe a 'level of inefficacy' for a model that is worse than the comparison baseline. Moreover, while d_r is piecewise continuous, it is not continuous through the second and higher derivatives. We explain why the coefficient of efficiency (E or E_2) and its modified form (E_1) are superior and preferable to many other statistics, including d_r, because of intuitive interpretability and because these indices have a fundamental meaning at zero.

We also expand on the discussion begun by Garrick *et al.* [Garrick M, Cunnane C, Nash JE. 1978. A criterion of efficiency for rainfall-runoff models. *Journal of Hydrology* **36**: 375-381.] and continued by Legates and McCabe [Legates DR, McCabe GJ. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**(1): 233-241.] and Schaefli and Gupta [Schaefli B, Gupta HV. 2007. Do Nash values have value? *Hydrological Processes* **21**: 2075-2080. DOI: 10.1002/hyp.6825.]. This important discussion focuses on the appropriate baseline comparison to use, and why the observed mean often may be an inadequate choice for model evaluation and development. Copyright © 2012 Royal Meteorological Society

KEY WORDS accuracy indices; coefficient of efficiency; goodness-of-fit; model evaluation; model-performance statistics

Received 15 November 2011; Accepted 12 March 2012

1. Introduction

A recent paper (Willmott *et al.*, 2012) examines various statistics that have been proposed and used in a wide variety of environmental fields to provide model evaluation. Model evaluation techniques are useful in model development, model verification, and model calibration, as well as in conveying model performance to others (Pappenberger and Beven, 2006; Schaefli and Gupta, 2007). In particular, Schaefli and Gupta (2007) note that while the Nash–Sutcliffe coefficient of efficiency (Nash and Sutcliffe, 1970) is widely used and well known by hydrologists, for example it is often foreign to researchers in other fields of environmental science. Moreover, values of the various statistics of model efficacy cannot be cross-compared and even those who often report such measures may not know what an index of model performance of 0.85 really means, for example.

Thus, Willmott *et al.*'s (2012) paper is vitally important in that it demonstrates a number of statistics of model evaluation available to researchers show little relationship across disparate measures. This underscores the need to understand more about each of these statistics and examine their behaviour and interpretation. However, the stated main purpose of Willmott *et al.* (2012) is the development of a 'refined' version of the dimensionless index of agreement (d_r) that resembles correlation in that it is bounded to the interval (-1,1]. Although they note that d_r is linearly equal to our modified coefficient of efficiency (E_1 – Legates and McCabe, 1999) over the positive portion of its domain, they contend that d_r represents an improvement over E_1. We respectfully disagree and argue that E_1 is superior owing to its more intuitive interpretation and other desirable characteristics.

*Correspondence to: D. R. Legates, Department of Geography, University of Delaware, Newark, DE 19716-2541, USA.
E-mail: legates@udel.edu.

2. The generic form of the coefficient of efficiency

Legates and McCabe (1999) wrote the generic form of the coefficient of efficiency, E_j, as

$$E_j = 1.0 - \frac{\sum_{i=1}^{N} |O_i - P_i|^j}{\sum_{i=1}^{N} |O_i - \overline{O}|^j} \qquad (1)$$

where the observed (O_i) and model predicted (P_i) series have N finite pairs for evaluation. In their original formulation, Nash and Sutcliffe (1970) used $j = 2$ (thus, $E \equiv E_2$), although Legates and McCabe suggested that $j = 1$ was a better scaling, owing to the fact that absolute values are preferable to squared terms since they do not give undue weight to outliers (see Willmott *et al.*, 1985; Willmott and Matsuura, 2005). The various coefficients of efficiency, E_js, are bounded by the range [1.0,$-\infty$) with a value of 1.0 indicating a perfect model (i.e. all $O_i = P_i$) and the statistics decrease as the model predicted and observed series diverge.

Of interest here is the interpretation of E_j. Although E_j has no lower bound, a value of $E_j = 0.0$ has a fundamental meaning. It implies that such a model has no more ability to predict the observed values than does the observed mean (or the baseline values, see Section 4.). In essence, since the model can explain no more of the variation in the observed values than can the observed mean, such a model can have no predictive advantage. For negative values of E_j, the model is less efficacious than the observed mean in predicting the variation in the observations. Although negative values of E_j represent a measure of the 'level of inefficacy' of a model (or maybe even a 'level of uselessness'), if you will, a negative value of E_j (or a negative 'Nash-Sutcliffe value', as they are called in hydrology) indicates that the model has failed to explain more of the variability in the observations than their mean.

Consider now the interpretation of E_2 (or E). The original formulation by Nash and Sutcliffe (1970) provides a direct comparison with the coefficient of determination, R^2, or the square of Pearson's product moment correlation coefficient. In simple linear regression, it is defined for the dependent variable, Y_i, as

$$R^2 = 1.0 - \frac{\left[\sum_{i=1}^{N} (Y_i - \hat{Y}_l)^2\right]}{\left[\sum_{i=1}^{N} (Y_i - \hat{Y})^2\right]} \qquad (2)$$

which is analogous to E_2 if $Y_i = O_i$, $\overline{Y} = \overline{O}$, and $\hat{Y}_l = P_i$ Thus, the interpretation of E_2 represents the percent of variance in the observations that is explained by the model predicted values. For example, a value of $E_2 =$ 0.75 implies that the model can explain three-quarters of the variance in the observed values. Like R^2, a value of $E_2 = 1.0$ implies a perfect model, whereas a value of $E_2 = 0.0$ implies a model that cannot explain any variance in the observations. Again, negative values of E_2 indicate that the model is worse than the observed mean (or baseline values) in predicting the observations.

Although E_1 (or other values of $j \neq 2$ in E_j) does not have this analogous property with R^2; we, nevertheless, believe that it is preferable to E_2 because of the decreased weight by E_1 on outliers. However, E_1 does provide a similar degree of interpretation, albeit with respect to absolute differences and not the variance (i.e. squared differences). A value of $E_1 = 0.75$, for example implies that the model is able to explain three-quarters of the absolute-valued differences between the observations and model predictions. Stated another way, the absolute value of the differences between the observations and the model predictions are only one-quarter of the difference between the observations and their mean (or baseline value). Such a model can explain 75% of the absolute difference between the observations and their mean.

This demonstrates that the scaling of the Nash–Sutcliffe coefficient of efficiency (E or E_2) and its modified form (E_1) is both robust and easily interpretable. We posit that this is a necessary quality for any metric of model evaluation. Such is why the coefficient of efficiency, in all its forms (i.e. E_j), is preferable to many other statistics available.

3. Limitations of the refined index of agreement

Willmott *et al.* (2012) define their refined index of agreement, d_r, as

$$d_r = \begin{cases} 1 - \dfrac{\sum_{i=1}^{N} |P_i - O_i|}{c\sum_{i=1}^{N} |O_i - \overline{O}|} \quad \text{when} \\ \sum_{i=1}^{N} |P_i - O_i| \leq c \sum_{i=1}^{N} |O_i - \overline{O}| \\ \dfrac{c\sum_{i=1}^{N} |O_i - \overline{O}|}{\sum_{i=1}^{N} |P_i - O_i|} - 1 \quad \text{when} \\ \sum_{i=1}^{N} |P_i - O_i| > c \sum_{i=1}^{N} |O_i - \overline{O}| \end{cases} \qquad (3)$$

where $c = 2.0$. They note that for $c = 1.0$, d_r is identically equal to the modified coefficient of efficiency, E_1, for the non-negative portion of its domain. Their refinement on E_1, therefore, is to rescale E_1 using c and remap the negative portion of d_r so it more resembles a correlation coefficient, that is d_r varies over the domain (-1.0, 1.0]. We argue that the choice of $c \neq 1.0$ destroys the interpretability of E_1, particularly at $E_1 = 0$ (and hence $d_r = 0$) and the remapping for negative values is merely cosmetic.

For $c = 2.0$, d_r attains a value of 0.0 when the absolute value of the difference between the model predicted values and observations equals twice the difference between the observations and their mean. That is, when $E_1 = 0.0$ $d_r = 0.5$ and when $d_r = 0.0$ $E_1 = -1.0$. Thus, a model with no more predictive ability than the observed mean would achieve a value of $d_r = 0.5$ and the value of $d_r = 0.0$ is arbitrary (i.e. the model has less than c^{-1} times the predictive ability of the observed mean). This makes the interpretation of d_r difficult and generally means that even relatively poor models will exhibit a high value of d_r, a criticism that also affects the original index of agreement (see Willmott *et al.*, 1985).

The remapping of the negative portion of d_r such that it scales from $(0,-\infty)$ to $(0,-1)$ is rather unnecessary. When $d_r = 0.5$, the model exhibits twice as much error (i.e. the absolute difference between the predicted and observed values) as could be achieved by using the observed mean as a predictor. In this sense, any value of $d_r < 0.5$ (as with a value of $E_1 < 0.0$) places the model as having an explanatory ability that is less than the observed mean. Thus, these values simply describe the 'level of inefficacy' of the model and it is largely immaterial how that portion of the statistic is scaled.

Moreover, although it is piecewise continuous, d_r exhibits a C_2 discontinuity (discontinuous second and higher derivatives) at 0.0. Let $A = \sum_{i=1}^{N} |P_i - O_i|$ and $B = \sum_{i=1}^{N} |O_i - \overline{O}|$ for simplicity. For the positive values of d_r (d_r^+),

$$\frac{d(d_r^+)}{dP_i} = -\frac{1}{cB} \text{ and } \frac{d^2(d_r^+)}{dP_i^2} = 0 \tag{4}$$

while for the negative values of d_r (d_r^-),

$$\frac{d(d_r^-)}{dP_i} = -\frac{cB}{A^2} \text{ and } \frac{d^2(d_r^-)}{dP_i^2} = 2\frac{cB}{A^3} \tag{5}$$

Thus, in the limit as $\sum|P_i - O_i| \to c\sum|O_i - \overline{O}_i|$ (i.e. $A \to cB$), d_r is continuous only through the first derivative.

4. Baseline adjustments

Willmott *et al.* (2012) provide a rather cursory paragraph on the issue of adjusted baselines. However, much commentary has been provided on this topic, and, as we feel it is an important consideration in model evaluation, it is necessary to examine this topic in more detail.

In general, all statistics for model evaluation compare the relative efficacy of the model to the predictive abilities of the observed mean. It forms the basis of evaluating regression performance using the coefficient of determination (R^2) and in the absence of any alternative model, the observed mean is the best 'strawman' against which a model can be compared. However, it has long been recognized that the observed mean may not be the best choice of a foil. For example, it was argued by Garrick *et al.* (1978, p. 376) that a comparison of a model to the observed mean was an 'unnecessarily primitive' choice. We demonstrated in Legates and McCabe (1999) that the evaluation of a model that predicts potential evapotranspiration in southern Louisiana or runoff in southwestern Colorado would lead to quite high values of any model evaluation statistic when compared against the observed mean. This is because any model that mimicked the seasonal cycle to a reasonable degree would significantly outperform the observed mean. As a result, Legates and McCabe (1999) proposed a further modification to our modified coefficient of efficiency, $E_1^{'}$, as

$$E_1^{'} = 1.0 - \frac{\sum_{i=1}^{N} |O_i - P_i|}{\sum_{i=1}^{N} |O_i - \overline{O}'_l|} \tag{6}$$

where the new baseline is $\overline{O}'_l$. Instead of simply comparing the observed values to a single number, such as the observed mean, the observed values can be compared against seasonally varying values, such as seasonal means, or a prediction using a function of other variables. The value of $\overline{O}'_l$ also could be used to define the results from a previous version of the model so that the statistic could represent the enhanced efficacy of the current model release. In any event, the choice of the observed mean can be easily substituted for a more appropriate baseline and researchers should be cognizant of the fact that the observed mean is not likely the most appropriate choice.

Schaefli and Gupta (2007) highlight the importance of specifying an appropriate baseline comparison and argue that it should become a standard practice in hydrologic modelling. We agree and argue that climatologists too must strongly consider comparing their models against more appropriate baselines. In particular, Schaefli and Gupta (2007, p. 2079) conclude,

> *Every modelling study should explain and justify the choice of benchmark. Of course, the appropriate benchmark will necessarily be different for different types of case studies. However, for efficient communication, the benchmark should fulfill the basic requirement that every [scientist] can immediately understand its explanatory power for the given case study and, therefore, appreciate how much better the actual [model] is.*

We wholly agree and argue that climatologists must consider model evaluation as an important component of their research and not simply as a statistic to be reported.

Such benchmarks are not likely to be globally applicable, but, as argued by Schaefli and Gupta (2007), we concur that it is important for each scientist to carefully select a benchmark that is appropriate for their particular study.

5. Concluding remarks

We are thankful that Willmott *et al.* (2012) have allowed us to extend the discussion from the hydrological community to climatological research. We wish to applaud the effort Dr Willmott has made in the statistical evaluation of model performance over the years. Moreover, we hope that such discussions lead climatologists in the future to take a more proactive role in model evaluation and the statistics that describe model efficiency.

We believe that our modified coefficient of efficiency (E_1) is an improvement over the Nash–Sutcliffe statistic (E_2) and forms the most appropriate basis for model evaluation based on its simplicity and ease of interpretation. The refined index of agreement, d_r, posited by Willmott *et al.* (2012) exhibits several distinct flaws that make its utility less favorable. However, we welcome a discussion regarding model evaluation and baseline comparisons that has begun in the hydrological sciences and hope it extends to the climatological community as well.

References

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

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

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

Pappenberger F, Beven KJ. 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research* **42**: W05302, DOI:10.1029/2005WR004820.

Schaefli B, Gupta HV. 2007. Do Nash values have value? *Hydrological Processes* **21**: 2075–2080, DOI: 10.1002/hyp.6825.

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, 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, Robeson SM, Matsuura K. 2012. A refined index of model performance. *International Journal of Climatology* forthcoming. DOI:10.1002/joc.2419.