

Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance

Author(s): Cort J. Willmott and Kenji Matsuura

Source: *Climate Research*, Vol. 30, No. 1 (December 19 2005), pp. 79-82

Published by: Inter-Research Science Center

Stable URL: <https://www.jstor.org/stable/10.2307/24869236>

REFERENCES

Linked references are available on JSTOR for this article:

https://www.jstor.org/stable/10.2307/24869236?seq=1&cid=pdf-reference#references_tab_contents

You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



Inter-Research Science Center is collaborating with JSTOR to digitize, preserve and extend access to *Climate Research*

JSTOR

NOTE

Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance

Cort J. Willmott*, Kenji Matsuura

Center for Climatic Research, Department of Geography, University of Delaware. Newark, Delaware 19716, USA

ABSTRACT: The relative abilities of 2, dimensioned statistics—the root-mean-square error (RMSE) and the mean absolute error (MAE)—to describe average model-performance error are examined. The RMSE is of special interest because it is widely reported in the climatic and environmental literature; nevertheless, it is an inappropriate and misinterpreted measure of average error. RMSE is inappropriate because it is a function of 3 characteristics of a set of errors, rather than of one (the average error). RMSE varies with the variability within the distribution of error magnitudes and with the square root of the number of errors ($n^{1/2}$), as well as with the average-error magnitude (MAE). Our findings indicate that MAE is a more natural measure of average error, and (unlike RMSE) is unambiguous. Dimensioned evaluations and inter-comparisons of average model-performance error, therefore, should be based on MAE.

KEY WORDS: Model-performance measures · Root-mean-square error · Mean absolute error

Resale or republication not permitted without written consent of the publisher

1. INTRODUCTION

Over the last few decades, there has been a proliferation in the number and types of climatic and environmental models. Interest also has increased in determining which formulations produce more accurate and precise estimates of the variables of interest; in turn, error statistics—which can be used to compare model-produced estimates with independent, reliable observations—have been applied more widely. Interest in evaluating differences among 2 or more comparable sets of estimates, when no set is known to be the most reliable, has grown as well, and this too has tended to increase the application of error or difference statistics. Recent papers (e.g. Fekete et al. 2004, Cavazos & Hewitson 2005) illustrate these trends.

Our purpose in this note is to explore and interpret available statistical measures of the average inaccuracy associated with a set of model-produced estimates.¹ More specifically, we reexamine the relative

abilities of 2, dimensioned measures of average model-performance error—the root-mean-square error (RMSE) and the mean absolute error (MAE). Each of these measures is ‘dimensioned’ in that it expresses average model-prediction error in the units of the variable of interest. These measures also have been used to represent average difference (rather than average error) when no set of estimates is known to be the most reliable. The RMSE is of particular interest because it is one of the most widely reported and misinterpreted error measures in the climatic and environmental literature.

¹Ideas expressed in this note emerged from efforts to refine and better apply model-performance statistics, especially in connection with the spatial estimation and evaluation of large-scale climate fields (e.g. see Willmott et al. 1985, Willmott & Matsuura 1995, Fekete et al. 2004). A preliminary analysis of several points developed here was presented at the 1996 annual meeting of the Association of American Geographers (Willmott et al. 1996)

*Email: willmott@udel.edu

© Inter-Research 2005 · www.int-res.com

2. DIMENSIONED MEASURES OF AVERAGE ERROR

Statistical comparisons of model estimates or predictions (P_i ; $i = 1, 2, \dots, n$) with thought-to-be reliable and pairwise matched observations (O_i ; $i = 1, 2, \dots, n$) remain among the most basic means of assessing model performance in the climatic and environmental sciences. Individual model-prediction errors usually are defined as $e_i = P_i - O_i$. Measures of average error or model performance then are based on statistical summaries of e_i ($i = 1, 2, \dots, n$).

Average model-estimation error can be written generically as

$$\bar{e}_\gamma = \left[\sum_{i=1}^n w_i |e_i|^\gamma / \sum_{i=1}^n w_i \right]^{1/\gamma} \quad (1)$$

where $\gamma \geq 1$ and w_i is a scaling assigned to each $|e_i|^\gamma$ according to its hypothesized influence on the total error (Willmott et al. 1985). Here, we let $w_i = 1.0$ for all i ; however, when i represents unequal areas and/or time intervals, such variability should be reflected in w_i . Average error is most commonly taken with $\gamma = 2$; that is, as the root-mean-square error (RMSE) where

$$\text{RMSE} = \left[n^{-1} \sum_{i=1}^n |e_i|^2 \right]^{1/2} \quad (2)$$

The stated rationale for squaring each e_i is usually 'to remove the sign' so that the 'magnitudes' of the errors influence the average error measure, RMSE. Considerably less often, average error has been assessed with $\gamma = 1$, according to

$$\text{MAE} = \left[n^{-1} \sum_{i=1}^n |e_i| \right] \quad (3)$$

which, of course, derives from the unaltered magnitude (absolute value) of each difference. This measure is commonly referred to as the mean absolute error or MAE. When $\gamma = 1$, but the signs of the errors are not removed, the average error becomes what is referred to as the mean bias error (MBE) or

$$\text{MBE} = n^{-1} \sum_{i=1}^n e_i = \bar{P} - \bar{O} \quad (4)$$

where \bar{P} and \bar{O} are the model-predicted and observed means, respectively. When MBE is reported (e.g. Fekete et al. 2004), it is usually intended to indicate average model 'bias'; that is, average over- or under-prediction. MBE can convey useful information, but should be interpreted cautiously since it is inconsistently related to typical-error magnitude, other than being an underestimate ($\text{MBE} \leq \text{MAE} \leq \text{RMSE}$). Since RMSE and, to a lesser extent, MAE are reported and interpreted in the literature, they are examined in more detail below.

3. ASSESSMENT OF MAE AND RMSE

Calculation of MAE is relatively simple. It involves summing the magnitudes (absolute values) of the errors to obtain the 'total error' and then dividing the total error by n ; once again, assuming that the w_i s are all equal to 1.0. Calculation of the RMSE involves a sequence of 3 simple steps. 'Total square error' is obtained first as the sum of the individual squared errors; that is, each error influences the total in proportion to its square, rather than its magnitude. Large errors, as a result, have a relatively greater influence on the total square error than do the smaller errors. This means that the total square error will grow as the total error is concentrated within a decreasing number of increasingly large individual errors. Total square error then is divided by n , which yields the mean-square error (MSE). The third and final step is to take RMSE as the square root of the MSE.

Since the division by n and the square root only scale the total square error, it follows that the MSE and RMSE will increase (along with the total square error) as the variance associated with the frequency distribution of error magnitudes increases. This often overlooked (and deleterious) property of MSE and RMSE is illustrated here with a simple, hypothetical example (Table 1): as the error-magnitude variance increases steadily—from Case 1 (where it is zero) through Case 5 (where it is at its maximum)—RMSE also increases steadily. It is apparent (in Table 1) that the lower limit of RMSE is MAE, which occurs only when $|e_1| = |e_2| = \dots = |e_n|$ or $e_1^2 = e_2^2 = \dots = e_n^2$. It is easily shown that the upper limit of RMSE is $n^{1/2} \cdot \text{MAE}$, which is reached when all of the prediction error is contained within a single error. The value of that single error is $n \cdot \text{MAE}$ and, since all of the other errors are zero, the total square error ($\sum_{i=1}^n |e_i|^2$) is $n^2 \cdot \text{MAE}^2$ and the upper limit of RMSE is $n^{1/2} \cdot \text{MAE}$. More concisely, the bounding of RMSE is

Table 1. Five hypothetical sets (cases) of 4 errors, and their corresponding totals, MAEs and RMSEs. Each e_i ($e_i = P_i - O_i$, $i = 1, 2, 3, 4$) is a hypothetical error value

Variable	Case 1	Case 2	Case 3	Case 4	Case 5
e_1	2	1	1	0	0
e_2	2	1	1	0	0
e_3	2	3	1	1	0
e_4	2	3	5	7	8
$\sum e_i $	8	8	8	8	8
MAE	2	2	2	2	2
$\sum e_i ^2$	16	20	28	50	64
RMSE	2.0	2.2	2.6	3.5	4.0

$MAE \leq RMSE \leq n^{1/2} \cdot MAE$. With the upper limit of RMSE increasing with $n^{1/2}$, while the lower limit is fixed at MAE, it also is true that RMSE generally increases with $n^{1/2}$. It is interesting to note that, when the errors are associated with grid-cell areas, n —or more precisely $\sum_{i=1}^n w_i$ —becomes the total area covered by the grid; similarly, when the errors are associated with consecutive time intervals, $\sum_{i=1}^n w_i$ becomes the length of time spanning the errors.

It is clear then that RMSE varies with the variability of the error magnitudes (or squared errors), as well as with the total-error or average-error magnitude (MAE) and $n^{1/2}$. Without benefit of other information (e.g. MAE), it is impossible to discern to what extent RMSE reflects central tendency (average error) and to what extent it represents the variability within the distribution of squared errors or $n^{1/2}$. A scatterplot of 10 pairs of MAE and RMSE (Fig. 1), calculated from differences among 5 mean-annual global precipitation fields, illustrates this partially. Each MAE/RMSE pair was drawn from a recent comparison of global precipitation data sets (Table 2 in Fekete et al. 2004). Each pair was calculated from the same difference field. It can be seen (Fig. 1) that RMSE is always larger than MAE, and that the difference between RMSE and MAE is increasing with MAE, but not monotonically. The inconsistency (scatter) between MAE and RMSE, in this instance, is only due to the differing error-magnitude variances associated with these sets of errors, since n , or the area covered by the grid, is constant. If n or the total area covered differed among these sets of errors, as it usually does when RMSEs from different studies are compared, the inconsistency could not be attributed to error-magnitude variance alone. In situations where it may be useful to know the variability within an error-magnitude field, we encourage researchers to compute and interpret a measure of error-magnitude variance, rather than to try to tease it out of RMSE.

Our view is that Fekete et al. (2004) should have presented MAE, and not RMSE, which was added in response to a reviewer's recommendation. As an unintended consequence, however, Fekete et al. now provide a good illustration (Fig. 1) of the inconsistent relationship between MAE and RMSE. Inconsistency is also apparent among the comparable MAEs and RMSEs more recently reported by Cavazos & Hewitson (2005). Such inconsistencies demonstrate our point that RMSE is not a true or reliable measure of 'average error.' Further, it should not be used to compare the average performance of 2 or more models, as is common practice (Fekete et al. 2004). This is especially true for RMSEs evaluated over different domains (different ns , areas, or time periods), because such RMSEs

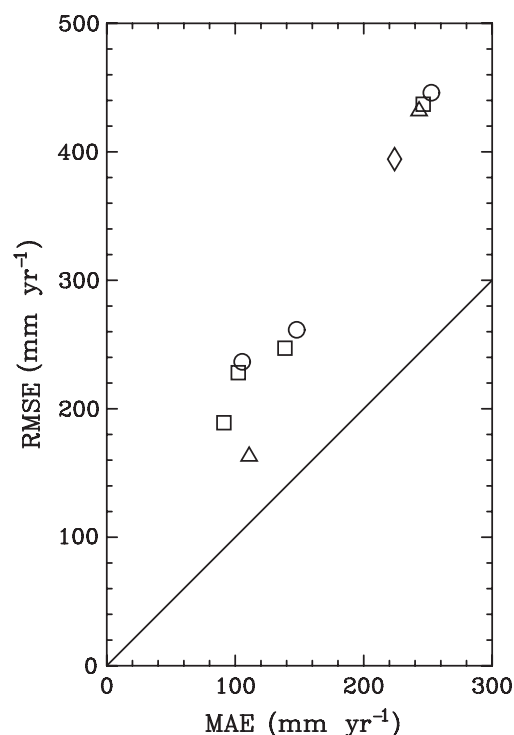


Fig. 1. MAE and RMSE values (mm yr^{-1}) associated with 10 combinations of the pairwise differences between 5 mean-annual, gridded terrestrial precipitation fields (see Table 2 in Fekete et al. 2004). When one land-surface precipitation field is subtracted from another, a difference is produced at each and every grid node. MAE and RMSE then are calculated from the gridded difference field. Data were made available by the Climatic Research Unit (CRU) of the University of East Anglia; Willmott and Matsuura (WM); the Global Precipitation Climate Center (GPCC); the Global Precipitation Climatology Project (GPCP); and the National Centers for Environmental Prediction, Atmospheric Model Intercomparison Project Reanalysis (NCEP2). When the same gridded mean-annual precipitation field is subtracted from the other mean-annual precipitation fields, the same plot symbol is used. (□) MAE/RMSE points associated with subtractions of the CRU field from the WM, GPCC, GPCP and NCEP2 mean-annual precipitation fields. (○) Average differences between the WM field, and the GPCC, GPCP and NCEP2 fields. (Δ) Average differences between the GPCC field, and the GPCP and NCEP2 fields. (◇) Average difference between the GPCP and NCEP2 fields

will vary with the square root of n (or of the area covered or the time domain) as well as with MAE and the variability within the set of error magnitudes. From a mathematical perspective, counter-intuitive values of RMSE are expected because $|e_i|^2$ (or e_i^2) is not a metric; that is, $|e_i|^2$ does not satisfy the triangle inequality of a metric (Mielke & Berry 2001). In sum, interpretation of RMSE is confounded because there is no consistent functional relationship between RMSE and average error.

Not only is it apparent that RMSE is an inappropriate indicator of average error, but it is also clear that $|e_i|$ is the actual size of each error and, in turn, the sum of all $|e_i|$ s (possibly weighted by the w_i s) is a meaningful measure of total error. The most natural measure of average error, therefore, must be MAE. Our view is consistent with earlier interpretations (e.g. Mielke 1985, Willmott et al. 1985, Willmott & Matsuura 1995, Mielke & Berry 2001, U.S. EPA 2003).

4. CONCLUDING REMARKS

Measures of average error (such as RMSE) that are based on the sum of squared errors (i.e. on the sum of e_i^2) are functions of the average error (MAE), the distribution of error magnitudes (or squared errors), and $n^{1/2}$; therefore, they do not describe average error alone. Among the disturbing characteristics of RMSE are: it tends to become increasingly larger than MAE (but not necessarily in a monotonic fashion) as the distribution of error magnitudes becomes more variable; and, it tends to grow larger than MAE with $n^{1/2}$, since its lower limit is fixed at MAE and its upper limit ($n^{1/2} \cdot \text{MAE}$) increases with $n^{1/2}$. For these reasons, it seems to us that there is no clear interpretation of RMSE or related measures, and we recommend that such measures no longer be reported in the literature. It also occurs to us that previous model-performance evaluations and inter-comparisons, which were based primarily on RMSE or related measures, are questionable and should be reconsidered. Other commonly used bivariate statistics that share RMSE's reliance on the sum of squares (e.g. certain correlation and skill measures) also are questionable model-performance statistics.

Our analysis indicates that MAE is the most natural measure of average error magnitude, and that (unlike RMSE) it is an unambiguous measure of average error magnitude. It seems to us that all dimensioned eval-

uations and inter-comparisons of average model-performance error should be based on MAE.

Acknowledgements. Several of the ideas presented in this paper are extensions of concepts previously considered by Willmott and his graduate students. In particular, we are indebted to D. R. Legates for his early recognition of the potential utility of a $\gamma = 1$ based version of Willmott's (1981) index of agreement. We also are indebted to P. W. Mielke, Jr. for his innovative work on distance-function statistics, and for his counsel. Four anonymous reviewers of this paper provided valuable interpretations, and we truly appreciate them. NASA's Seasonal to Interannual Earth Science Information Partner (SIESIP), and the NSF Arctic RIMS project at the University of New Hampshire (NSF OPP 0230243) funded portions of this research.

LITERATURE CITED

- Cavazos T, Hewitson BC (2005) Performance of NCEP-NCAR reanalysis variables in statistical downscaling of daily precipitation. *Clim Res* 28:95–107
- Fekete BM, Vörösmarty CJ, Roads JO, Willmott CJ (2004) Uncertainties in precipitation and their impacts on runoff estimates. *J Clim* 17:294–304
- Mielke PW Jr (1985) Geometric concerns pertaining to applications of statistical tests in the atmospheric sciences. *J Atmos Sci* 42:1209–1212
- Mielke PW Jr, Berry KJ (2001) *Permutation methods: a distance function approach*, Springer-Verlag, New York
- U.S. EPA (2003) Technical support document for the Clear Skies Act 2003: air quality modeling analyses. Environmental Protection Agency, Office of Air Quality Planning and Standards, Emissions Analysis and Monitoring Division, Research Triangle Park, NC
- Willmott CJ, Matsuura K (1995) Smart interpolation of annually averaged air temperature in the United States. *J Appl Meteorol* 34:2577–2586
- 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. *J Geophys Res* 90(C5):8995–9005
- Willmott CJ, Webber SR, Power HC (1996) Better statistics for the evaluation of model performance. Annual meeting of the Association of American Geographers (AAG), Charlotte, North Carolina. Abstract published in the 1996 AAG Annual Meeting Abstracts. Washington, DC: AAG 316–317

Editorial responsibility: Robert E. Davis, Charlottesville, Virginia, USA

*Submitted: April 4, 2005; Accepted: September 13, 2005
Proofs received from author(s): December 15, 2005*