Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature

Tianfeng Chai[1,2] **and Roland R. Draxler**[1]

[1]NOAA Air Resources Laboratory (ARL), NOAA Center for Weather and Climate Prediction, 5830 University Research Court, College Park, MD 20740, USA
[2]Cooperative Institute for Climate and Satellites, University of Maryland, College Park, MD 20740, USA

Correspondence to: T. Chai
(Tianfeng.Chai@noaa.gov)

Abstract

Both the root mean square error (RMSE) and the mean absolute error (MAE) are regularly employed in model evaluation studies. Willmott and Matsuura (2005) have suggested that the RMSE is not a good indicator of average model performance and might
5   be a misleading indicator of average error and thus the MAE would be a better metric for that purpose. While some concerns over using RMSE raised by Willmott and Matsuura (2005) and Willmott et al. (2009) are valid, the proposed avoidance of RMSE in favor of MAE is not the solution. Citing the aforementioned papers, many researchers chose MAE over RMSE to present their model evaluation statistics when presenting
10 or adding the RMSE measures could be more beneficial. In this technical note, we demonstrate that the RMSE is not ambiguous in its meaning, contrary to what was claimed by Willmott et al. (2009). The RMSE is more appropriate to represent model performance than the MAE when the error distribution is expected to be Gaussian. In addition, we show that the RMSE satisfies the triangle inequality requirement for a dis-
15 tance metric whereas Willmott et al. (2009) indicated that the sums-of-squares-based statistics do not satisfy this rule. In the end, we discussed some circumstances where using the RMSE will be more beneficial. However, we do not contend that the RMSE is superior over the MAE. Instead, a combination of metrics, including but certainly not limited to RMSEs and MAEs, are often required to assess model performance.

20 **1   Introduction**

The root mean square error (RMSE) has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies. The mean absolute error (MAE) is another useful measure widely used in model evaluations. While they have both been used to assess model performance for many
25 years, there is no consensus on the most appropriate metric for model errors. In the field of geosciences, many present the RMSE as a standard metric for model errors

(e.g. McKeen et al. (2005); Savage et al. (2013); Chai et al. (2013)), while a few others choose to avoid the RMSE and present only the MAE, citing the ambiguity of the RMSE claimed by Willmott and Matsuura (2005) and Willmott et al. (2009) (e.g. Taylor et al. (2013); Chatterjee et al. (2013); Jerez et al. (2013)). When both metrics are calculated, the MAE tends to be much smaller than the RMSE because the RMSE penalizes large errors while the MAE gives the same weight to all errors. For instance, Chai et al. (2009) presented both the mean errors (MAE) and the RMS error (RMSE) of model NO_2 column predictions compared to SCIAMACHY satellite observations. The ratio of RMSE to MAE ranged from 1.63 to 2.29 (see Table 1 of Chai et al. (2009)).

Using hypothetical sets of 4 errors, Willmott and Matsuura (2005) demonstrated that while keeping the MAE as a constant of 2.0, the RMSE varies from 2.0 to 4.0. They concluded that the RMSE varies with the variability of the the error magnitudes and the total-error or average-error magnitude (MAE), and the sample size n. They further demonstrated an inconsistency between MAEs and RMSEs using 10 combinations of 5 pairs of global precipitation data. They summarized that the RMSE tends to become increasingly larger than the MAE (but not necessarily in a monotonic fashion) as the distribution of error magnitudes becomes more variable. The RMSE tends to grow larger than the MAE with $n^{\frac{1}{2}}$ since its lower limit is fixed at the MAE and its upper limit ($n^{\frac{1}{2}} \cdot$ MAE) increases with $n^{\frac{1}{2}}$. Further, Willmott et al. (2009) concluded that the sums-of-squares-based error statistics such as the RMSE and the standard error have inherent ambiguities and recommended the use of alternates such as the MAE.

As every statistical measure condenses a large number of data into a single value, it only provides one projection of the model errors emphasizing a certain aspect of the error characteristics of the model performance. Willmott and Matsuura (2005) have simply proved that the RMSE is not equivalent to the MAE and one cannot easily derive the MAE value from the RMSE (and vice versa). Similarly, one can readily show that for several sets of errors with the same RMSE, the MAE would vary from set to set.

Since statistics are just a collection of tools, researchers must select the most appropriate tool for the question being addressed. Because the RMSE and the MAE are

defined differently, we should expect the results to be different. Sometimes multiple metrics are required to provide a complete picture of error distribution. When the error distribution is expected to be Gaussian and there are enough samples, the RMSE has an advantage over the MAE to illustrate the error distribution.

[5]   The objective of this note is to clarify the interpretation of the RMSE and the MAE. In addition, we demonstrate that the RMSE satisfies the triangle inequality requirement for a distance metric, whereas Willmott and Matsuura (2005) and Willmott et al. (2009) have claimed otherwise.

2   Interpretation of RMSE and MAE

[10] To simplify, we assume that we already have n samples of model errors ϵ calculated as (e_i, i=1,2,...,n). The uncertainties brought in by observation errors or the method used to compare model and observations are not considered here. We also assume the error sample set ϵ is unbiased. The RMSE and the MAE are calculated for the data set as

[15]

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2} \tag{2}$$

     The underlying assumption when presenting the RMSE is that the errors are un-
[20] biased and follow a normal distribution. Thus, using the RMSE or the standard error (SE)[1] helps to provide a complete picture of the error distribution.

[1]For unbiased error distributions, the standard error (SE) is equivalent to the RMSE as the sample mean is assumed to be zero. For an unknown error distribution, the SE of mean is the

Table 1 shows RMSEs and MAEs for randomly generated pseudo-errors with zero mean and unit variance Gaussian distribution. When the sample size reaches 100 or above, using the calculated RMSEs one can re-construct the error distribution close to its "truth" or "exact solution", with its standard deviation within 5 % to its truth, i.e. $SE = 1$. When there are more samples, reconstructing the error distribution using RMSEs will be even more reliable. The MAE here is the mean of the half-normal distribution, i.e., the average of the positive subset of a population of normally distributed errors with zero mean. Table 1 shows that the MAEs converge to 0.8, an approximation to the expectation of $\sqrt{\frac{2}{\pi}}$. It should be noted that all statistics are less useful when there are only a limited number of error samples. For instance, Table 1 shows that neither the RMSEs nor the MAEs are robust when only 4 or 10 samples are used to calculate those values. In those cases, presenting the values of the errors themselves, e.g. in tables, are probably more appropriate than calculating any of the statistics. Fortunately, there are often hundreds of observations available to calculate model statistics, unlike the examples with n=4 (Willmott and Matsuura, 2005) and n=10 (Willmott et al., 2009).

Condensing a set of error values into a single number, either the RMSE or the MAE, removes a lot of information. The best statistics metrics should not only provide a performance measure, but also a representation of the error distribution. The MAE is suitable to describe uniformly distributed errors. Because model errors are likely to have a normal distribution rather than a uniform distribution, the RMSE is a better metric to present than the MAE for such type of data.

---

square root of the "bias-corrected sample variance". That is, $SE = \sqrt{\frac{1}{n-1} \sum\limits_{i=1}^{n} (e_i - \bar{\epsilon})^2}$, where $\bar{\epsilon} = \frac{1}{n} \sum\limits_{i=1}^{n} e_i$.

3 Triangle inequality of a metric

Both Willmott and Matsuura (2005) and Willmott et al. (2009) emphasized that sums-of-squares-based statistics do not satisfy the triangle inequality. An example is given in a footnote of Willmott et al. (2009). In the example, it is given that d(a,c)=4, d(a,b)=2, and
5   d(b,c)=3, where d(x,y) is a distance function. The authors stated that d(x,y) as a "metric" should satisfy the "triangle inequality", i.e. $d(a,c) \leq d(a,b) + d(b,c)$. However, they did not specify what a, b, and c represent here before arguing that the sum of squared errors does not satisfy the "triangle inequality" because $4 \leq 2+3$ whereas $4^2 \not\leq 2^2 + 3^2$. In fact, this example represents the mean-square-error (MSE) which cannot be used
10 as a distance metric, rather than the RMSE.

Following a certain order, the errors e_i, i=1,...,n can be written into a n-dimensional vector ϵ. The L1-norm and L2-norm are closely related to the MAE and the RMSE, respectively, as shown in Equations 3 and 4.

$$15 \quad |\boldsymbol{\epsilon}|_1 = (\sum_{i=1}^{n} |e_i|) = n \cdot MAE \tag{3}$$

$$|\boldsymbol{\epsilon}|_2 = \sqrt{(\sum_{i=1}^{n} e_i^2)} = \sqrt{n} \cdot RMSE \tag{4}$$

All vector norms satisfy $|\boldsymbol{X} + \boldsymbol{Y}| \leq |\boldsymbol{X}| + |\boldsymbol{Y}|$ and $|-\boldsymbol{X}| = |\boldsymbol{X}|$ (see, e.g. Horn and Johnson, 1990). It is trivial to prove that the distance between two vectors measured
20 by Lp-norm would satisfy $|\boldsymbol{X} - \boldsymbol{Y}|_p \leq |\boldsymbol{X}|_p + |\boldsymbol{Y}|_p$. With three n-dimensional vectors, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, we have

$$|\boldsymbol{X} - \boldsymbol{Y}|_p = |(\boldsymbol{X} - \boldsymbol{Z}) - (\boldsymbol{Y} - \boldsymbol{Z})|_p \leq |\boldsymbol{X} - \boldsymbol{Z}|_p + |\boldsymbol{Y} - \boldsymbol{Z}|_p. \tag{5}$$

For n-dimensional vectors and the L2-norm, Equation 5 can be written as

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \leq \sqrt{\sum_{i=1}^{n}(x_i - z_i)^2} + \sqrt{\sum_{i=1}^{n}(y_i - z_i)^2}, \tag{6}$$

which is equivalent to

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - z_i)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - z_i)^2}. \tag{7}$$

5 This proves that RMSE satisfies the triangle inequality required for a distance function metric.

4 Summary and discussion

We present that the RMSE is not ambiguous in its meaning, and is more appropriate to use than the MAE when model errors follow a normal distribution. In addition, we
10 demonstrate that the RMSE satisfies the triangle inequality required for a distance function metric.

The sensitivity of the RMSE to outliers is the most common concern with the use of this metric. In fact, the existence of outliers and their probability of occurrence is well described by the normal distribution underlying the use of the RMSE. Table 1
15 shows that with enough samples (n≥100), including those outliers, one can closely re-construct the error distribution. In practice, it might be justifiable to throw out the outliers that are several orders larger than the other samples when calculating the RMSE, especially if the number of samples is limited. If the model biases are severe, one may also need to remove the systematic errors before calculating the RMSEs.
20 One distinct advantage of RMSEs over MAEs is that RMSEs avoid the use of absolute value which is highly undesirable in many mathematical calculations. For instance,

it might be difficult to calculate the gradient or sensitivity of the MAEs with respect to certain model parameters. Furthermore, in the data assimilation field, the sum of squared errors is often defined as the cost function to be minimized by adjusting model parameters. In such applications, penalizing large errors through the defined least-square terms proves to be very effective in improving model performance. Under the circumstances of calculating model error sensitivities or data assimilation applications, MAEs are definitely not preferred over RMSEs.

An important aspect of the error metrics used for model evaluations is their capability to discriminate among model results. The more discriminating measure that produces higher variations in its model performance metric among different sets of model results is often the more desirable. In this regard, the MAE might be affected by a large amount of average error values without adequately reflecting some large errors. Giving higher weighting to the unfavorable conditions, the RMSE usually is better at revealing model performance differences. In many of the model sensitivity studies that use only RMSE, a detailed interpretation is not critical because variations of the same model will have similar error distributions. When evaluating different models using a single metric, differences in the error distributions become more important. As we stated in the note, the underlying assumption when presenting the RMSE is that the errors are unbiased and follow a normal distribution. For other kinds of distributions, more statistical moments of model errors, such as mean, variance, skewness, and flatness, are needed to provide a complete picture of the model error variation. Some approaches that emphasize resistance to outliers or insensitivity to non-normal distributions have been explored by other researchers (Tukey, 1977; Huber and Ronchetti, 2009).

As stated earlier, any single metric provides only one projection of the model errors, and therefore only emphasizes a certain aspect of the error characteristics. A combination of metrics, including but certainly not limited to RMSEs and MAEs, are often required to assess model performance.

References

Chai, T., Carmichael, G. R., Tang, Y., Sandu, A., Heckel, A., Richter, A., and Burrows, J. P.: Regional NO$_x$ emission inversion through a four-dimensional variational approach using SCIA-MACHY tropospheric NO$_2$ column observations, Atmos. Environ., 43, 5046–5055, 2009.

Chai, T., Kim, H. C., Lee, P., Tong, D., Pan, L., Tang, Y., Huang, J., McQueen, J., Tsidulko, M., and Stajner, I.: Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO$_2$ measurements, Geosci. Model Dev., 6, 1831–1850, 2013.

Chatterjee, A., Engelen, R. J., Kawa, S. R., Sweeney, C., and Michalak, A. M.: Background error covariance estimation for atmospheric CO$_2$ data assimilation, J. of Geophys. Res., 118, 10 140–10 154, 2013.

Horn, R. A. and Johnson, C. R.: Matrix Analysis, Cambridge University Press, 1990.

Huber, P. and Ronchetti, E.: Robust statistics, Wiley New York, 2009.

Jerez, S., Pedro Montavez, J., Jimenez-Guerrero, P., Jose Gomez-Navarro, J., Lorente-Plazas, R., and Zorita, E.: A multi-physics ensemble of present-day climate regional simulations over the Iberian Peninsula, Climate Dynamics, 40, 3023–3046, 2013.

McKeen, S. A., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Hsie, E., Gong, W., Bouchet, V., Menard, S., Moffet, R., McHenry, J., McQueen, J., Tang, Y., Carmichael, G. R., Pagowski, M., Chan, A., Dye, T., Frost, G., Lee, P., and Mathur, R.: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, J. of Geophys. Res., 110, Art. No. D21 307, 2005.

Savage, N. H., Agnew, P., Davis, L. S., Ordóñez, C., Thorpe, R., Johnson, C. E., O'Connor, F. M., and Dalvi, M.: Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation, Geosci. Model Dev., 6, 353–372, 2013.

Taylor, M. H., Losch, M., Wenzel, M., and Schroeter, J.: On the Sensitivity of Field Reconstruction and Prediction Using Empirical Orthogonal Functions Derived from Gappy Data, J. of Climate, 26, 9194–9205, 2013.

Tukey, J. W.: Exploratory Data Analysis, Addison-Wesley, 1977.

Willmott, C. and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res., 30, 79–82, 2005.

5 Willmott, C. J., Matsuura, K., and Robeson, S. M.: Ambiguities inherent in sums-of-squares-based error statistics, Atmos. Environ., 43, 749–752, 2009.

Table 1. RMSEs and MAEs of random generated pseudo-errors with a zero mean and unit variance Gaussian distribution. Five sets of errors of size n are generated with different random seeds.

n	RMSEs	MAEs
4	0.92, 0.65, 1.48, 1.02, 0.79	0.70, 0.57, 1.33, 1.16, 0.76
10	0.81, 1.10, 0.83, 0.95, 1.01	0.65, 0.89, 0.72, 0.84, 0.78
100	1.05, 1.03, 1.03, 1.00, 1.04	0.82, 0.81, 0.79, 0.78, 0.78
1000	1.04, 0.98, 1.01, 1.00, 1.00	0.82, 0.78, 0.80, 0.80, 0.81
10,000	1.00, 0.98, 1.01, 1.00, 1.00	0.79, 0.79, 0.79, 0.81, 0.80
100,000	1.00, 1.00, 1.00, 1.00, 1.00	0.80, 0.80, 0.80, 0.80, 0.80
1000,000	1.00, 1.00, 1.00, 1.00, 1.00	0.80, 0.80, 0.80, 0.80, 0.80