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

Tianfeng Chai, Roland Draxler

Response to Reviewers' comments

May 1, 2014

We are very grateful to the reviewers for reading the manuscript carefully and forwarding their valuable comments. Point-by-point responses to the reviewers' comments and concerns are listed below.

Response to Anonymous Referee #1:

General Comments:

This is the third time that I have been asked to review a version of this paper, twice previously for the journal Atmospheric Environment which ultimately elected not to publish this paper. The authors of "Root mean square error (RMSE) or mean absolute error (MAE)?" argue that Willmott and collaborators. recommendations.to preferentially use the MAE, rather than the RMSE, as the appropriate measure of average error magnitude in model performance evaluations (articulated in papers published in 2005 and 2009) is "not the solution to the problem". They claim to "demonstrate that the RMSE is not ambiguous in its meaning and [it] is more appropriate to represent model performance than the MAE when the error distribution is expected to be Gaussian". They also report "that the RMSE satisfies the triangle inequality requirement".

Besides the two similar reviews we had from the same referee, the other review was very positive, stating "The manuscript is well-structured and well-written, and can certainly be considered for publication in Atmospheric Environment". Instead of going through the repeal process, we chose the open discussion provided by this journal to give readers an opportunity to evaluate arguments from both sides. We still think Willmott and Matsuura (2005) and Willmott et al (2009) might mislead "uncritical researchers" to avoid the RMSE based upon unsubstantial arguments. As the other reviewer stated, one should consider the error distribution before deciding which metric to use.

Specific Comments:

- *This manuscript falls well short of refuting Willmott et al.'s main points about the interpretational limitations associated with the RMSE; because its value is a function of three variables (the MAE, the variability among the error magnitudes and, in instances, the square root of n). Even though the authors claim to show that the RMSE is not ambiguous, they do not. They appear unable to appreciate that, if one does not know what portion of an RMSE is average error magnitude and what portion is error magnitude variability, it (an RMSE) really is "ambiguous" and impossible to adequately interpret or meaningfully compare with other RMSEs. Without additional information, all that can be said about the RMSE is that it inconsistently overestimates average error magnitude. And, if one knows the values of the above-mentioned variables that comprise an RMSE, it (the RMSE) provides no additional information and is superfluous. If, in addition to the average error magnitude (the MAE), variability among the error magnitudes is of interest, a measure of error magnitude variability should be calculated, reported and interpreted in addition to the MAE, as suggested by Willmott et al. (2009). The MAE, in contrast to the RMSE, is a straightforward measure of average error magnitude.*

  Explaining the RMSE as a function of three variables (the MAE, the variability among the errors, and the square root of n) is not needed to interpret the RMSE. 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, the RMSE provides the first order information on its variability. Higher order information on the error variability, such as skewness and flatness, can be obtained using higher moments of the error distribution.

  We do not agree that the RMSE is "ambiguous" only because the MAE value cannot be easily derived from the RMSE. In fact, there is no ambiguity for any metric with a clear definition and a mathematical formula. The interpretation provided in our note is straightforward, or "rather basic undergraduate textbook material", as the other reviewer put it.

- *The authors want to assume errors are normally distributed and unbiased, and this seems to be their primary argument for evaluating the RMSE. It's difficult to understand why anyone would want to assume normality when they do not have to.*

  In the note, we drew samples from a normal distribution just to show that the RMSE can be easily interpreted. In model evaluation applications, people do not need to assume normality as one can easily check the actual error distribution. When the errors are available for either the RMSE or the MAE calculation, the distribution can

be easily plotted as well. For most applications, it is not unusual to have 100 or more error samples. Furthermore, the Gaussian distribution is probably the most common distribution of the model errors. The error distributions that we plotted using our own data or we saw in various publications are often close to Gaussian (a log-normal distribution is ubiquitous as well).

- *Nonetheless, the authors say that "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". This is a red herring, for at least three reasons. First, their example (see their Table 1) clearly shows that the MAE does equally well at recovering their ("Gaussian") error distribution. Second, whether their preconditions are satisfied adequately is nearly always questionable. Third, regardless of whether these preconditions satisfied, one still should know the average error magnitude (the MAE), which cannot be teased out of the RMSE post hoc.*

  First, using the MAE to recover the Gaussian distribution is clearly counter-intuitive. Second, as stated earlier, whether the error distribution is close to a normal distribution can be easily checked when the errors are already available for the RMSE or the MAE calculation. Third, we do not object using the MAEs. Sometimes, it is necessary to present both the RMSE and the MAE. As we stated in the note, any 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.

- *In addition, let me mention that "modern statistics" increasingly tends to emphasize interpretability over "exact properties". Subfields such as "exploratory data analysis" and "robust statistics" have produced a number of innovative approaches that enhance interpretability (e.g., through resistance to outliers or insensitivity to non-normal distributions) more than mathematical exactness. Classic texts by Tukey (1977) and Huber (1981), for example, introduce some intriguing approaches and I encourage the authors to explore such more flexible options.*

  We thank the reviewer for providing the references to this interesting topic. This has been mentioned in our note while citing Tukey (1977) and Huber and Ronchetti (2009, the 2nd edition).

- *Even though the triangular inequality issue is a relatively minor one, it is unfortunate that the authors of this paper misinterpret Willmott et al.'s correct point about*

combinations of elements within the squared-errors vector not satisfying the triangular inequality. Although Willmott et al. explained their invocation of the triangular inequality; perhaps their brief descriptions of this (in both 2005 and 2009) were insufficient, because the authors of this manuscript have misinterpreted them. In the 2005 paper, for instance, Willmott and Matsuura state that "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 the 2009 paper, Willmott et al. similarly indicate that "the relative influence of each squared error on the sum of the squared errors often is counter-intuitive [in part, because of likely triangular inequalities among the individually squared errors], which undermines meaningful scientific interpretation of the sum of the squared errors" and, in turn, of the RMSE. In other words, relationships among the squared errors (within the set of squared errors from which the RMSE is evaluated) that do not satisfy the triangular inequality, when the squared errors are summed, can help produce nonsensical sums of squares and RMSEs.

We agree that the mean-square-error (MSE) cannot be used as a distance metric, but how this fact undermines meaningful interpretation of the RMSE is not clear. In fact, the abstract of Willmott et al (2009) states, "Interpretation problems arise, according to Paul Mielke, because **sums-of-squares-based statistics do not satisfy the triangle inequality**". RMSE is given as an example of the "sums-of-squares-based statistics" in the same abstract, where the first sentence reads, "Commonly used sums-of-squares-based error or deviation statistics - like the standard deviation, the standard error, the coefficient of variation, and the root-mean-square error ... ". In the note, we clearly showed that RMSE and MAE are closely related to L1-norm and L2-norm for vectors and they both satisfy the triangular inequality. It is necessary to clarify that the triangular inequality should be applied to the metric [RMSE] itself, rather than to the separate terms [each squared errors] that construct the metric.

We did thoroughly consult Mielke and Berry (2001) again and found there are fundamental differences between their context and the RMSE metric issue discussed here. Mieke and Berry (2001) focuses on the application of multiresponse permutation procedure (MRPP). In permutation tests, pairs or groups generated from a set of objects are not totally independent of each other. For instance, 9 objects in MRPP can generate $\binom{9}{2} = 36$ different sample pairs. In such a context, adding up 36 squared Euclidean distances calculated from the sample pairs is not expected to generate sensible properties, as demonstrated in the book. However, in the applications where the "MAE

or RMSE" question arises, no permutation tests are needed or performed. Errors are calculated using observations and model results. The independency among the calculated errors are often well justified. Then, adding the squared errors together on the way to get RMSE is equivalent to adding squared velocity components (in Euclidean coordinates) in order to get the wind speed.

Response to Anonymous Referee #2:

- *Overall, I was a bit surprised to find that this is a debatable topic. First, trying to argue that either MAE or RMSE is superior seems like a strange motivation for a paper. The former should be used when the data or model error is known (or suspected) to follow a Laplacian distribution. The latter should be used when the data or model error is known (or suspected) to follow a Gaussian distribution. This has been known for a long time (centuries?). So demonstrating that samples drawn from one distribution do not fit the other, while true, does not advance the state of knowledge regarding error statistics. Second, the demonstration of what is, or is not, a proper metric, is also rather basic undergraduate textbook material.*

  We had the same opinion as the reviewer, i.e. neither MAE nor RMSE is superior over the other and one should consider the error distribution before deciding which metric to use. We did not feel that was debatable or needed to be clarified any further until we read the two papers, Willmott and Matsuura (2005) and Willmott et al (2009) which strongly recommend avoiding RMSE in favor of MAE. As a result, many cited the aforementioned papers when choosing MAE over RMSE in presenting their evaluation results.

  "The basic undergraduate textbook material" shows that RMSE is most appropriate "when model error is known (or suspected) to follow a Gaussian distribution" and it is not ambiguous at all, counter to the arguments made by Willmott and Matsuura (2005). It also shows that the RMSE satisfies the triangle inequality of a distance metric just as the MAE does. Although these discussions do not "advance the state of knowledge regarding error statistics", they provide an opportunity for other researchers to more carefully consider the statistical approaches they use in model evaluations.

- *Overall, I do not feel this note warrants publication because the concepts being presented are very simplistic (definition of a metric), and the perspective is a bit incomplete in that demonstrating that Gaussian errors do not fit a Laplacian distribution, or vice versa, is a stawman argument, neglecting the more important point of correctly matching the metric used for evaluation with one's best estimate of the statistics. That being said, if there are indeed cases in the literature where this has been confused, it is perhaps worth pointing out. So I would encourage the authors to include their arguments as a sub-point within an actual research paper, and perhaps even demonstrate the consequences of mismatching the appropriate metric with the statistics of the model / data error on a forecast evaluation; it just doesn.t seem to me like a substantial enough point*

to make on its own.

It is true that the material presented is very simplistic and does not warrant publication as a regular research paper. However, as the reviewer stated, it is worth pointing out the confusion in the literature. Including the content as a sub-point in a larger research paper, as the reviewer suggested, risks getting this topic unnoticed with respect to the main point of the paper and thus not as effective as having a standalone short technical note. We added a subtitle, "- Arguments against avoiding RMSE in the literature", to clarify the motivation of the paper.

Response to Short Comment by James Lee :

Both RMSE and MAE have been widely used in model performance evaluation. After reading this article and the article by Willmott et al., 2009, I would think this article further enhance our understanding on the statistic factors used in numerical model evaluations. The publication of this article can improve the use of these statistic items in Geosciences. I did not find any possible mistakes in this article. The math and logicality in this article is accurate and consistent. Therefore, I favor a final publication after addressing a series of issues that can improve the expression on this topic and may improve the feeling of readers on this article.

Thanks for your short comment. Below are our responses to the two specific points.

1. *Both this article and the article by Willmott et al., 2009 did not strictly and clearly define the problems and impact regime. The article by Willmott et al., 2009 may have problem in using samples to derive its conclusion, so that the conclusion on RMSE is loaded too broad. My experience in using RMSE really have better functions in evaluation of weather models than that of MAE, especially for extreme events simulations. This article did not directly point the possible problem in Willmott et al., 2009. Therefore, I suggest further revision on this part to provide a clear analysis of possible problems that may affect the misleading conclusion in the article by Willmott et al., 2009.*

   The conclusion made by Willmott and Matsuura (2005) and Willmott et al., 2009 to avoid RMSE are mainly based on the following two arguments.

   - The RMSE does not directly provide the average error magnitude information, i.e. the MAE.
   - "Interpretation problems arise, according to Paul Mielke, because sums-of-squares-based statistics do not satisfy the triangle inequality".

   For the first point, we have shown in text that it is not necessary to derive the MAE from the RMSE value. Just like all other metrics, the RMSE and the MAE are defined differently, without ambiguity, to provide the unique statistics they are designed to provide. For the second point, we prove that the RMSE and the MAE are closely related to L1-norm and L2-norm for vectors and they both satisfy the triangular inequality.

   The reviewer commented that, "My experience in using RMSE really have better functions in evaluation of weather models than that of MAE, especially for extreme

events simulations". It can be explained by the better discriminating aspect of RMSE over MAE. Discussion on this has been added in "Summary and discussion".

2. *Some parts of expressions and statements in this article may need to be revised, so that the valuable part of previous works can be better cited. This include using soft language in commenting the problems by the target article (Willmott et al., 2009), and convince the valuable part of previous studies.*

   We have modified the manuscript and changed some expressions. Specifically, the abstract has been rewritten. We also acknowledged some concerns raised by Willmott et al. (2009) when using the RMSEs.