Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC1
https://doi.org/10.5194/gmd-2022-64-AC1, 2022

Reply on RC1

Timothy Hodson

Author comment on "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not" by Timothy O. Hodson, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-64-AC1, 2022

Thank you taking time to review my manuscript. You make several good points that have helped me to clarify and strengthen my arguments. I don't agree with some of them, but they are all important to consider and illustrate the confusion left by earlier papers on this topic.

Major points

1. Regarding the abstract:
Yes, the abstract is reductive, particularly in the case of Chai and Draxler, but their paper is somewhat inconsistent in its arguments. They do state clearly their belief that neither metric is inherently better but offer little explanation for why and, instead, list several reasons for preferring RMSE to MAE. For the abstract, I would be willing to frame Chai and Draxler in this manner, rather than as favoring RMSE.

Chai and Draxler (2014) state their objective as "to clarify the interpretation of the RMSE and the MAE." I agree that this is an incredibly important topic, but I also believe their paper has important flaws. Rather than providing a point-by-point rebuttal to their work, my paper focuses on the classic proofs for why and when RMSE and MAE work. Besides settling some aspects of the debate, these proofs prepare the reader to understand how formal likelihoods can address the limitations of RMSE and MAE.

In responding to RC1, I will present some of that point-by-point rebuttal, focusing on three of Chai and Draxler's arguments (listed in their order of occurence):

Argument 1: "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 shows that with enough samples (n = 100), including those outliers, one can closely re-construct the error distribution."

Argument 2: "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 a type of data."

Argument 3: "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."

2. Regarding my warning against using multiple metrics:

The reviewer argues that in most applications, evidence to support one metric over the other is not easily attainable. This is not so. The law of likelihood states the evidence for one metric versus another is simply the likelihood ratio. I've shown how to compute the likelihoods associated with MAE and RMSE (the Laplace and normal, respectively). The "evidence" is simply the ratio of the two, which is easily attainable. In practice, one must also adjust for differences in degrees of freedom (yielding the AIC), which is described in detail in Burnham and Anderson (B&A). I cited B&A, but I will add a statement to this effect.
As the likelihood ratio approaches unity, it is reasonable to consider multiple metrics weighted by their "evidence."
I agree with Argument 3 that each metric presents a different measure (transformation) of the error, but there is an infinite variety of such transformations. How do we arrive at the best one? Why not error to the fourth power, etc? The standard approach is to select several candidates based on prior knowledge of the system, then weight them by the evidence (B&A). I discuss both steps in the paper, though I neglect to describe how they are used in conjunction. I will briefly mention that.

Specific points

1. Abstract Lines 3-4:
The abstract is reductive, but abstracts have some license to be so. Chai and Draxler (2014) do state that neither metric is inherently better (Argument 3); however, they go on to list several arguments for why they prefer RMSE to MAE (Arguments 1,2, and others). These two sides are never completely reconciled in their paper. Many of their arguments for favoring RMSE are flawed, and beyond Argument 3, they offer none of the theory underlying their claim that "neither metric is better," only a simulation, which they use to simultaneously claim RMSE is better, and neither is better, ultimately advocating that it's best to present both metrics, as well as others.

Consider Arguments 1 and 2. In Argument 1, they simulate a normal with a standard deviation of 1, then verify that the standard deviation of the result is 1. While this confirms the RMSE is appropriate for normals, it gives little explanation of why this is so. They go on to claim this as evidence that RMSE is robust to outliers, thereby suggesting that MAE, which is often preferred for its robustness, is superfluous. But the "outliers" in the normal distribution are not the "outliers" of robust statistics, which deals with fatter-tailed distributions or other deviations from the normal that are common in practice. Argument 2 is similarly unclear. It seems to say that MAE is suited only for uniform distributions, which are atypical, so RMSE is better, while simultaneously saying that RMSE is only better for normal distributions.

2. Lines 33-34 "That recent shift may explain why Willmott and Matsuura (2005) and Chai and Draxler (2014) were unaware of the historical justification for MAE and RMSE; neither were they the first to overlook it":

The reviewer suggests I omit this opinion. I am willing to do so, but neither paper references the proofs, which would have negated this debate. Chai and Draxler were aware that RMSE assumes a normal error distribution, though they do not describe or reference why. However, they seem unaware of the classic proof for MAE, or else why would they associate MAE with uniform case, a case which they later claim is irrelevant

(Argument 2), rather than focusing on the Laplacian or the contaminated normal, for which MAE is optimal and superior, respectively?

My intent with this comment was to remind the reader that the RMSE-MAE debate has come up several times before. I do not wish to offend the other authors. Prestigeous scientists, including R.A. Fischer, made a similar oversight, and I believe reminding readers of this fact is important context. I would also like to be clear that the intended target of my critique is not individual authors but a longstanding failure by the Earth-science community to fully integrate probability theory into its modeling practices. I'd prefer to make that claim directly, but I think it would draw additional criticism. In the interest of keeping this paper short and instructive, I defer that debate.

3. Equation 10: revised

4. Line 180: revised

5. Line 209 suggesting that Chai and Draxler argue that MAE only applies to uniform errors:
The reviewer is correct that Chai and Draxler don't explicitly make this claim, but I'm probably not alone in interpreting them this way: see Argument 2 for example, nor do they describe a case for which MAE would be better suited other than the uniform distribution, which they dismiss as not being useful. They mention the classic argument that MAE works better in the presence of outliers, but dismiss this as well. After claiming the sensitivity of RMSE is not a practical concern (Argument 1), they go on to say "in practice, it might be justifiable to throw out the outliers that are several orders larger than the other samples when calculating the RMSE." The first statement seems to contradicts the second, which admits the concern is well warranted. Why do Chai and Draxler argue for throwing out data points, rather than acknowledging this a case where MAE (or MAD) may be better? I cannot answer say, but it is another example of the dissonance within their paper.