Author's response for "Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not"

Timothy O. Hodson[1]

[1]U.S. Geological Survey Central Midwest Water Science Center, Urbana, IL

Correspondence: Timothy O. Hodson (thodson@usgs.gov)

1 Responses to RC1

1.1 RC1: Comment 1

1.1.1 Comment

The abstract mostly states the motivation of the paper and a not well-substantiated opinion. The reader might need to rewrite the abstract to better reflect the actual content of this paper.

1.1.2 Response

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. I will revise the abstract to reflect this.

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 the comments by 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 occurrence):

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.”

1.1.3 Revision

1.2 RC1: Comment 2

1.2.1 Comment

In Section 6 of the paper, “Why not use both RMSE and MAE?,” the author argues against using both RMSE and MAE. It is stated that presenting both metrics is “unnecessary and potentially confusing” “If the evidence strongly supports one over the other.” In most applications, such evidences to strongly support one metric over the other are not easily attainable. While decomposing one metric into several independent components suggested by the author is viable, using multiple metrics is still a practical way to avoid mistaken conclusions caused by merely relying on one metric.

1.2.2 Response

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.

1.2.3 Revision

Assuming no prior information, the logical approach is to weigh them (the metrics) by their likelihoods. According to the law of likelihoods, the evidence for one hypothesis versus another corresponds to the ratio of their likelihoods (Edwards, 1992, , page 30). Extending this further, either metric can be weighted based on its relative likelihood (Burnham and Anderson, 2001).

If the evidence strongly supports one over the other, presenting both metrics is unnecessary and potentially confusing. If their evidence is similar, than it may be appropriate to present a weighted average (Burnham and Anderson, 2001). When averaging informal likelihoods to estimate the typical error, an additional adjustment must be made for differences in their scale, as demonstrated with MAD. Priors can be incorporated as well, though this is a more advanced topic.

1.3 RC1: Comment 3

1.3.1 Comment

Abstract Lines 3-4, “Some of this confusion arises from a recent debate between Willmott and Matsuura (2005) and Chai and Draxler (2014), in which either side presents their arguments for one metric over the other.”

While Chai and Draxler (2014) argued against favoring MAE over RMSE by Willmott and Matsuura (2005), they did not favor RMSE over MAE. That is clearly stated in their abstract.

1.3.2 Response

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.

1.3.3 Revision

See the revised abstract, which states that either author presents “arguments for favoring one metric or the other.”

1.4 RC1: Comment 4

1.4.1 Comment

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 author might be entitled to have such a judgment of others’ unawareness or overlook, but it is better to avoid such opinions in a scientific paper.

1.4.2 Response

I will omit this statement. Later in the text I will say that neither author gives the historical justification (proofs) for MAE or RMSE.

1.4.3 Revision

Deleted.

1.5 RC1: Comment 5

1.5.1 Comment

Equation 10: A close parenthesis is missing

1.5.2 Response

Revised.

1.5.3 Revision

1.6 RC1: Comment 6

1.6.1 Comment

Line 180: Please add a comma after “Laplace”.

1.6.2 Response

Revised.

1.7 RC1: Comment 7

1.7.1 Comment

Line 209. “. . . though wrongly suggest that MAE only applies to uniformly distributed errors”:

It is not accurate. Although Chai and Draxler (2014) gave one example with “uniformly distributed errors” to show that the MAE would be a good metric for such cases, it was not suggested that they are the ONLY cases where MAE would be appropriate. This statement is a misinterpretation of the paper.

1.7.2 Response

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.

1.7.3 Revision

Since it is unclear what Chai and Draxler intended, I will revise this as "though they seem to suggest, wrongly, that MAE only applies to uniformly distributed errors."

2 Response to RC2

2.1 RC2: Comment 1

2.1.1 Comment

I question the suggestion that there is a debate between Willmott's papers and Chai and Draxler (2014). In my mind, a debate has some back and forth and here we only see a comment by Chai & Draxler, but nothing in response from Willmott. Perhaps an alternate word would be more precise.

2.1.2 Response

I will frame the debate as between RMSE and MAE, in which Chai and Willmott are a recent installment.

2.1.3 Revision

See revised abstract: "In a recent reprise to the two-century-long debate over their use, Willmott and Matsuura (2005) and Chai and Draxler (2014) present arguments for favoring one metric or the other."

2.2 RC2: Comment 2

2.2.1 Comment

While I found the paper interesting, I returned several times considering the question "Who is the audience?" I think that presenting in a broader fashion for a wider audience would widen its appeal. Probably it is just me being 'old school' but the use of "we" is sometimes confusing as it is not specific that it is referring to only the author, or the wider community.

2.2.2 Response

My audience are readers of Willmott and Chai (>3000 citations each): Earth scientists with limited statistical training who use least squares (MSE) and least absolute deviations (MAE) extensively but have little-to-no awareness of formal likelihood methods. To paraphrase Burnham and Anderson (2001) in comparing lleast squares to likelihood methods, "likelihood methods are much more general and far less taught." I attempted to write a brief pedagogical paper that describes how the familiar terms like RMSE and MAE arise from likelihood theory, gives examples of likelihoods' greater generality, then refers the reader to the important textbooks on this topic. The discussion of MAD is somewhat tangential to likelihood methods, but robustness is a common point in the "debate" between RMSE and MAE, and MAD is relevant in that respect.

2.2.3 Revision

Added a note that "the review is primarly written for Earth scientists who use RMSE or MAE but have little-to-no awareness of formal likelihood methods."

2.3 RC2: Comment 3

2.3.1 Comment

What is the metric for? How does it need to be applied? How does this affect how my model can be used? Too often, the metrics are simply a checklist and little thought seems to be applied to questioning whether the model is fit for purpose. How does any metric help address "Do you get the right answer for the right reasons?"

2.3.2 Response

The task is simple, in theory: choose the metric that will identify the most likely ("realistic") model; for normal iid errors, minimizing the MSE yields the most likely model. I agree with the reviewer that too often we blindly apply "checklists." This paper introduces a more theory-based approach to evaluation, which has long been the standard other fields like ecology and economics.

2.4 RC2: Comment 4

2.4.1 Comment

Another area that needs mentioning is that models are fit to data; data that often is assumed to be without error.

2.4.2 Response

The short answer is that, likelihoods are easily extended to accommodate observational error. The longer answer is that in practice we frequently don't know the exact nature of the observational errors, but even a rough model is better than no model. In either case, assuming we're comparing models with the same observational uncertainty, those uncertainties cancel when computing the likelihood ratio. This is an important subject, but it's more relevant to the related topic of assessing accuracy/confidence limits and less so in model selection.

2.5 RC2: Comment 5

2.5.1 Comment

One area that needs clarification is the application to models. The manuscript never clearly suggests the modeling framework intended; the key issue is often that the observations and model output are time series and the residuals are unlikely to be iid but strongly autocorrelated. This is particularly true since the hydrological examples are for rainfall-runoff modeling. But, the arguments regarding MAE and MSE apply to random sampling as well.

2.5.2 Response

I suppose the main framework would be "likelihood methods," though not exclusively. All rational frameworks (Bayesian, significance testing, etc) can be derived from probability theory. A point of the paper, is to remind readers that MAE versus MSE is a false dichotomy and whatever framework they choose, they should understand how it derives from probability theory. Autocorrelation is an important topic that could be addressed within the likelihood framework, but may be a bit advanced. Like Chai and Willmott, I sought to write a short readable paper but also to orient readers to the existing literature. I reference several papers that discuss autocorrelation in the context of rainfall-runoff modeling, though I consider this an advanced topic.

2.6 RC2: Comment 6

2.6.1 Comment

Line 15 "L1-norm and L2-norm" would be better to explain that L1-norm is Manhattan distance and L2-norm is Euclidean distance. Could explain more fully here.

2.6.2 Response

I'll note that, but I'd prefer to omit the equations. The conceptual link is important, but the equations are tangential here, I think.

2.7 RC2: Comment 7

2.7.1 Comment

Line 18 is it actually a "debate"?

2.7.2 Response

Would "discussion" or "discourse" be better I'm not sure. According to one source, a debate is "a formal discussion on a particular topic in a public meeting or legislative assembly, in which opposing arguments are put forward." I believe that definition is consistent with Willmott and Chai, but I will try to frame the 'debate' as the two-century-long debate, in which Willmott and Chai are a recent iteration.

2.7.3 Revision

See revised abstract, but I've replaced this with "discussion," which is a more neutral term.

2.8 RC2: Comment 8

2.8.1 Comment

Line 27. I like the "historical" presentation.

2.8.2 Response

Thank you.

2.9 RC2: Comment 9

2.9.1 Comment

Line 31. Would be good to add a bit more guidance.

2.9.2 Response

I'm uncertain what sort of guidance was intended. This line describes how many reference works give the proofs behind MSE and MAE but neglect to give a primary reference.

2.10 RC2: Comment 10

2.10.1 Comment

Line 37 insert after observations “and models”

2.10.2 Response

Revised

2.11 RC2: Comment 11

2.11.1 Comment

Line 37 for choosing between MAE and RMSE

2.11.2 Response

Revised

2.12 RC2: Comment 12

2.12.1 Comment

Line 38 “for the complex error . . . ”

2.12.2 Response

Revised.

2.13 RC2: Comment 13

2.13.1 Comment

Line 39 I would choose a better word than “Occasionally” “Where more concrete examples were needed, the examples were drawn from hydrology, particularly rainfall-runoff modeling.”

2.13.2 Response

Revised

2.14 RC2: Comment 14

2.14.1 Comment

Line 43 delete “In their debate,”

2.14.2 Response

Revised

2.15 RC2: Comment 15

2.15.1 Comment

Line 59 delete “Despite its simplicity,”

2.15.2 Response

Revised.

2.16 RC2: Comment 16

2.16.1 Comment

Line 61 “simpler problem of deduction” “harder problem of induction” Not sure that these value words help as they take a stance that is not necessary.

2.16.2 Response

I believe this is an accurate summary of the frequentist argument.

2.17 RC2: Comment 17

2.17.1 Comment

Line 72 “under certain conditions” It would be better to explain those conditions.

2.17.2 Response

The next sentence transitions into a more detailed discussion, which states these conditions.

2.18 RC2: Comment 18

2.18.1 Comment

Line 81. Replace “Subsequent sections will . . . ” with something like “Ways of relaxing these assumptions will be introduced below.” [The subsequent text covers much more that what was indicated here.]

2.18.2 Response

Revised.

2.19 RC2: Comment 19

2.19.1 Comment

Line 90 Choose a better word than “trick.” Perhaps “operation”? There was a case in recent memory where emails from East Anglia referred to a “trick” that the media ceased as evidence of subterfuge and dishonesty.

2.19.2 Response

Good point. Revising as a "convenient practice."

2.20 RC2: Comment 20

2.20.1 Comment

Line 125-127 seems awkward. The text regarding using likelihood functions to inform choices of model structure seems tangential.

2.20.2 Response

Revised.

2.20.3 Revision

The preceding derivations were formulated in terms of maximizing the likelihood by way of adjusting the model parameters θ, but more generally, the likelihood can be used to refine the entire model—both its parameters and structure.

2.21 RC2: Comment 21

2.21.1 Comment

Line 127 insert a period after “. . . 2001)”

2.21.2 Response

Revised.

2.22 RC2: Comment 22

2.22.1 Comment

Line 132 “often approximately log normal” and it would be good to specify that the underlying assumption is for only perennial streams.

2.22.2 Response

Revised.

2.23 RC2: Comment 23

2.23.1 Comment

Line 135 sentence requires citing a reference.

2.23.2 Response

I’ve found several papers that claim this, but none of them are primary. I think the point is that if we’re comfortable thinking in linear or proportional scales, but it’s more difficult to reason about other nonlinear scales.

2.24 RC2: Comment 24

2.24.1 Comment

Also, here the converse is the real problem: interpreting the “results” without the transformation. While the units would be “correct” the model assumptions would not be. It is also possible to transform, analyze, and retransform so the units are “correct” but you face asymmetric confidence limits.

2.24.2 Response

The confidence limits are symmetric in geometric space; the error is multiplicative rather than additive Limpert et al. (2001, see).

2.25 RC2: Comment 25

2.25.1 Comment

Line 139 “Student’s-t”

2.25.2 Response

Revised.

2.26 RC2: Comment 26

2.26.1 Comment

Line 144 use “normal distribution” rather than “normal condition.”

2.26.2 Response

Revised.

2.27 RC2: Comment 27

2.27.1 Comment

Line 145 Perhaps the text regarding Tukey’s contributions in general is tangential?

2.27.2 Response

True, Tukey was seminal but also intermediary. I did include several historical tangents lest we repeat them. Many people use MAE as a robust metric because of Tukey’s work, including Willmott, so I wanted to place Tukey in context. However, the primary purpose of the paper is pedagogical, so I’d considering omitting this comment if it is distracting.

2.28 RC2: Comment 28

2.28.1 Comment

Line 148 “Neither option is ideal” or “Neither option is acceptable”?

2.28.2 Response

I prefer ideal (or optimal). I want to advocate that these are better practices, but I don’t want to go so far as to claim that all others are unacceptable, though some arguably are (by “better”, I mean more efficient, more accurate, etc.).

2.29 RC2: Comment 29

2.29.1 Comment

Line 150 “Since Tukey’s work, some alternatives have emerged.” “better” seems to be a convenient opinion.

2.29.2 Response

Better in respect to the preceding statement “their performance degrades as deviation grows.” Will revise as “more robust”

2.30 RC2: Comment 30

2.30.1 Comment

Line 156 “RMSE is inefficient (or more inefficient) for error. . . ”

2.30.2 Response

Revised.

2.31 RC2: Comment 31

2.31.1 Comment

“Lacking an alternative, MAD is a popular choice.”

2.31.2 Response

Revised.

2.32 RC2: Comment 32

2.32.1 Comment

“Log transforming”

2.32.2 Response

Revised.

2.33 RC2: Comment 33

2.33.1 Comment

Line 165 Not sure I would agree that this provides “reasonable” results. There are methods for dealing with the zero elements and non-zero elements separately.

2.33.2 Response

Those methods are discussed later, though you’re right that “reasonable” was a poor choice as it implies they are “logical” or “rational.” Describing these methods as “practicable” or “sometimes satisfactory” would be better.

2.33.3 Revision

Simple work arounds, such as setting zeros to a small positive value, may be satisfactory when zero and near-zero values are relatively rare but blow up as those values become more frequent.

2.34 RC2: Comment 34

2.34.1 Comment

Line 167 “… so the difference between 0.001 and 1 is same as that between 1 and 1000.

2.34.2 Response

Revised

2.35 RC2: Comment 35

2.35.1 Comment

Line 172 “Log-likelihood is equivalent to the concept of entropy from information theory.” While true, it seems tangential to the main argument and distracts the reader.

2.35.2 Response

Will omit.

2.36 RC2: Comment 36

2.36.1 Comment

Line 180 “… normal versus Laplace; Burnham and Anderson …”

2.36.2 Response

Revised.

2.37 RC2: Comment 37

2.37.1 Comment

Line 189 use “non-zero” instead of ‘positive’

2.37.2 Response

This sentence is describing the zero-inflated lognormal, so it is strictly positive. True, streamflow can take negative values, so lognormal isn’t ideal, but that is an open problem. My intent is only to demonstrate how formal likelihoods offer additional flexibility beyond MSE, which assumes errors are normal and equal variance.

2.38 RC2: Comment 38

2.38.1 Comment

Line 195 chose a word other than “naively” ... “without considerable thought”

2.38.2 Response

Revised

2.39 RC2: Comment 39

2.39.1 Comment

Line 200 combining metrics is certainly not meaningful. Neither is a checklist of metrics without criteria or benchmarks. This is a place where the issue of “how can I use my model?” follows. What do the metrics tell you about the limits of model applicability?

2.39.2 Response

I’ll add some brief guidance here. I focused on methods for determining the most likely model. Assessing confidence or applicability is a related topic (via probability theory), but I hadn’t planned on addressing it in this paper.

2.39.3 Revision

This review has focused primarly on how probability theory can answer the question “which model is better?”, thereby guiding the task of model selection. But this task is essentially the same as asking “how accurate is my model?”, then comparing the

accuracy of competing models. The first step—quantify the uncertainty in our models—is import in its own right epecially if we intend to base decisions on predictions from our models. Just as the Gaussian likelihood provides the theoretical basis for using RMSE to quantify the typical error among normally distributed errors, more generally, other likelihood functions provide the basis for assessing the accuracy and confidence intervals of models with other error distributions, though that subject is beyond the scope of this review.

2.40 RC2: Comment 40

2.40.1 Comment

Line 201 replace “best” with “most important with respect to model application.”

2.40.2 Response

Will simply say “while exposing readers to several alternatives for when they fail.”

Disclaimer. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Burnham, K. P. and Anderson, D. R.: Kullback-Leibler information as a basis for strong inference in ecological studies, Wildlife Research, 28, 111, https://doi.org/10.1071/wr99107, 2001.

Chai, T. and Draxler, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geoscientific Model Development, 7, 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014, 2014.

Edwards, A. W. F.: Likelihood: Expanded edition, Johns Hopkins University Press, 1992.

Limpert, E., Stahel, W. A., and Abbt, M.: Log-normal Distributions across the Sciences: Keys and Clues, BioScience, 51, 341, https://doi.org/10.1641/0006-3568(2001)051[0341:lndats]2.0.co;2, 2001.

Willmott, C. J. and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Climate Research, 30, 79–82, https://doi.org/10.3354/cr030079, 2005.