Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC3
https://doi.org/10.5194/gmd-2022-64-AC3, 2022

Reply on RC2

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-AC3, 2022

Thank you taking time to review my manuscript. Your comments were insightful and I will address all of them, except one that I didn't understand, which I note in my response.

Major Comments

RC2: 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.

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

RC2: 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.

AC: 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 Burhnam and Anderson's (2001) comparison of least 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.

RC2: On the use of value laden words [e.g. better, best, simpler, and harder]:

AC: Good point. I'll find better descriptors.

RC2: 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?"

AC: 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.

RC2: Another area that needs mentioning is that models are fit to data; data that often is assumed to be without error.

AC: Observational error is a common application of likelihood and Bayesian methods. I could mention some relevant texts, though I'm less familiar with the history.

RC2: 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.

AC: 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. I will think about a better general reference that I could cite.

Specific Comments

RC2: 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.
AC: 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.

RC2: Line 18 is it actually a 'debate'?
AC: 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 consistent with Willmott and Chai. I will also try to frame the 'debate' as the two-century-long debate, in which Willmott and Chai are a recent iteration.

RC2: Line 27. I like the "historical" presentation.
AC: Thank you.

RC2: Line 31. Would be good to add a bit more guidance.
AC: 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.

RC2: Line 37 insert after observations “and models”
AC: revised

RC2: Line 37 for choosing between MAE and RMSE
AC: revised

RC2: Line 38 “for the complex error …”
AC: revised

RC2: 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.”
AC: revised

RC2: Line 43 delete “In their debate,”
AC: revised

RC2: Line 59 delete “Despite its simplicity,”
AC: revising as "This simple equation provides..."

There are several places in the text where value laden words as used loosely.

RC2: 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.
AC: I believe this is an accurate summary of the frequentist argument. I try not to take a strong stance on this subject, but I would say I'm Bayesian in theory but sometimes frequentist in practice.

RC2: Line 72 “under certain conditions” It would be better to explain those conditions.
AC: The next sentence transitions into a more detailed discussion, which states these conditions.

RC2: 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.]
AC: revised

RC2: 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.
AC: Good point. This is a "convenient practice."

RC2: Line 125-127 seems awkward. The text regarding using likelihood functions to inform choices of model structure seems tangential.
AC: revised

RC2: Line 127 insert a period after “… 2001)”
AC: revised

RC2: Line 132 “often approximately log normal” and it would be good to specify that the underlying assumption is for only perennial streams.
AC: revised

RC2: Line 135 sentence requires citing a reference.

AC: 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.

RC2: 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.
AC: The confidence limits are symmetric in geometric space; the error is multiplicative rather than additive. See Limpert et al. (2001): Log-normal distributions across the sciences.

RC2: Line 139 "Student's-t"
AC: revised

RC2: Line 144 use "normal distribution" rather than "normal condition"
AC: revised

RC2: Line 145 Perhaps the text regarding Tukey's contributions in general is tangential?
AC: 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.

RC2: Line 148 "Neither option is ideal" "Neither option is acceptable"?
AC: 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.).

RC2: Line 150 "Since Tukey's work, some alternatives have emerged." "better" seems to be a convenient opinion.
AC: Better in respect to the preceding statement "their performance degrades as deviation grows." Will revise as "more robust"

RC2: Line 156 "RMSE is inefficient (or more inefficient) for error…"
AC: revised

RC2: "Lacking an alternative, MAD is a popular choice."
AC: revised

RC2: Line 164 "Log transforming"
AC: revised

RC2: 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.
AC: 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. XXX TODO

RC2: Line 167 "… so the difference between 0.001 and 1 is same as that between 1 and 1000.
AC: revised

RC2: 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.

AC: Will omit.

RC2: Line 180 “… normal versus Laplace; Burnham and Anderson …”
AC: revised

RC2: Line 189 use “non-zero” instead of ‘positive’
AC: 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.

RC2: Line 195 chose a word other than “naively” “without considerable thought”
AC: revised

RC2: 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?
AC: 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.

RC2: Line 201 replace “best” with “most important with respect to model application.”
AC: Will simply say "while exposing readers to several alternatives for when they fail"