Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2022-64-RC2, 2022

Comment on gmd-2022-64

Anonymous Referee #2

Referee 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-RC2, 2022

Title: Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not

Author(s): Timothy O. Hodson

MS No.: gmd-2022-64

MS type: Review and perspective paper

This is a generally interesting and nicely written paper. Nice to see a presentation that looks to the source material; these days, too many people cite their own, their friends, and derivative work. Would be nice to see more authors placing their work in a more correct perspective.

I think it is useful to thoroughly explain reasoning as is done through much of the paper; there are places where more explanation could be provided. Some have been indicated in the detailed comments provided below.

I question the suggestion that there is a debate between Wilmott'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.

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. Similarly, there seems to be unnecessary use of value laden words without sufficient support and discussion [e.g. better, best, simpler, and harder]. Many of these could be avoided and the presentation would benefit from logical explanation and support rather that an implied author opinion.

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?"

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

One area that needs clarification is the application to models. The manuscript never clearly suggests the modelling 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 modelling. But, the arguments regarding MAE and MSE apply to random sampling as well.

Detailed comments:

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.

Line 18 is it actually a 'debate'?

Line 27. I like the "historical" presentation.

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

Line 37 insert after observations "and models"

Line 37 for choosing between MAE and RMSE

Line 38 “for the complex error …”

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 modelling.”

Line 43 delete “In their debate,”

[[but there are observations, models, and theory/understanding

Line 59 delete “Despite its simplicity,”

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

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.

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

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.]

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.

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

Line 127 insert a period after "… 2001)"

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

Line 135 sentence requires citing a reference.

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, analyse, and retransform so the units are ‘correct’ but you face asymmetric confidence limits.

Line 139 “Student’s-t”

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

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

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

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

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

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

Line 164 “Log transforming”

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.

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

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.

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

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

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

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?

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