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)

Abstract. The ~~mean absolute error (MAE) and~~ root mean squared error (~~RMSE~~RMSE) and mean absolute error (MAE) are widely used metrics for evaluating models. Yet, there remains enduring confusion over their use, such that a standard practice is to present both, leaving it to the reader to decide. ~~Some of this confusion arises from a recent debate between~~ In a recent reprise to the two-century-long debate over their use, Willmott and Matsuura (2005) and Chai and Draxler (2014) ~~, in which either side presents their arguments for one metric over~~ give arguments for favoring one metric or the other. ~~Neither side was completely correct; however, because neither~~ But this comparison can present a false dichotomy. Neither metric is inherently better: ~~MAE~~ RMSE is optimal for ~~Laplacian~~ normal (Gaussian) errors, and ~~RMSE~~ MAE is optimal for ~~normal (Gaussian)~~ Laplacian errors. When errors deviate from these distributions, other metrics are superior.

1 Introduction

The ~~mean absolute error (MAE) and~~ root mean squared error (~~RMSE~~RMSE) and mean absolute error (MAE) are two standard metrics used in model evaluation. For a sample of n observations y (y_i, $i = 1, 2, \ldots, n$), and n corresponding model predictions $\hat{y}$ the MAE and RMSE are

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2}$$

(3)

As its name implies, the RMSE is the square root of the mean squared error (MSE). Taking the root does not affect the relative ranks of models, but it yields a metric with the same units as y, which conveniently represents the typical or "standard" error for normally distributed errors. The ~~MAE and MSE~~ MSE and MAE are averaged forms of the ~~L1-norm and~~ L2-norm and L1-norm, which are ~~more common forms in math and statistics~~the Euclidean and Manhattan distance, respectively.

In what have become two classic papers in the geoscientific modeling literature, Willmott and Matsuura (2005, MAE) and Chai and Draxler (2014, RMSE) ~~debate whether MAE or RMSE~~ discuss whether RMSE or MAE is superior. In their introduction, Chai and Draxler (2014) state:

> The RMSE has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies. The MAE is another useful measure widely used in model evaluation. While they have both been used to assess model performance for many years, there is no consensus on the most appropriate metric for models errors.

That statement may have accurately characterized the geosciences but not statistics. Among statisticians, the answer was common knowledge, at least to the extent that there can be no consensus. Different types of models have different error distributions and so necessitate different error metrics. In fact, the debate over squared versus absolute error terms had emerged, was subsequently forgotten, and reemerged over the preceding two centuries (Boscovich, 1757; Gauss, 1816; Laplace, 1818; Eddington, 1914; Fisher, 1920), with history given by (Stigler, 1973, 1984), making it one of the oldest questions in statistics.

It is unclear exactly when this "no-solution solution" became common knowledge, in part, because contemporary authors rarely cite ~~these~~ its sources. While reviewing the literature, I found proofs in several reference works, including the venerable Press et al. (1992, p. 701), but no references to primary literature.

~~With the growth of machine learning, probability theory is experiencing a renaissance among the broader scientific community. 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.~~ As this review will show, the choice of error term should conform with the expected probability distribution of the errors~~. Small mismatches are frequently encountered in practice and can have large consequences on inference~~ ; otherwise, inference will be biased. The choice ~~among~~ of error metric is, therefore, fundamental in determining what scientists learn from their observations and models. This paper reviews the basic justification for choosing ~~MAE or RMSE~~ between RMSE or MAE, then discusses several alternatives better suited for the complex error distributions that are encountered in practice. The literature on this topic is vast, and I try to emphasize classic papers and textbooks from the statistical literature. ~~Occasionally, more concreteexamples are desirable, which I draw~~ To make that discussion more concrete, I include several examples from hydrology and rainfall-runoff modeling, though none of the techniques are exclusive to that field. The discussion is primarily written for Earth scientists who use RMSE or MAE but have little-to-no awareness of formal likelihood methods.

2 The naive (frequentist) basis

~~In their debate,~~ Willmott and Matsuura (2005) and Chai and Draxler (2014) present several arguments for and against MAE and RMSE. I will not review them here~~, but instead focus~~ . Instead, my focus is on an important ~~flaw: neither side explains that~~ omission in these papers: neither explains the theoretical justification for MSE and MAE~~are not arbitrary formulas. Rather, both metrics derive directly~~ , which derives from the laws of probability, which themselves derive from the laws of logic (Jaynes, 2003); that is to say, there are logical reasons for choosing one over the other.

Like all inference problems, the justification begins with Bayes' theorem,

$$\overbrace{p(\theta|y)}^{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{p(y)} \tag{4}$$

where y is some set of observations, θ are the model parameters, and $p(\theta|y)$ is the probability of θ given y. In words, Bayes' theorem represents the logical way of using observations to update our understanding of the world. The numerator of the right-hand side contains two terms: the prior, representing our state of knowledge before observing y; and the likelihood, representing what was learned by observing y. The left-hand side, known as the posterior, represents our updated state of knowledge after the observation. Given a set of observations y, the denominator of the right-hand side is constant, so, for convenience, Bayes' theorem is often rewritten as the proportion between the posterior and the product of the likelihood with the prior,

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{5}$$

In the absence of any prior information the prior distribution $p(\theta)$ is "flat" or constant, such that the posterior is simply proportional to the likelihood,

$$p(\theta|y) \propto p(y|\theta). \tag{6}$$

~~Despite its simplicity, the previous equation~~ This relation provides the basis for "frequentist" statistics, first recognized by Bernoulli (1713) and later popularized by Karl Pearson, Ronald Fisher, and others. Criticisms of frequentism aside (see Clayton, 2021, for summary), the recognition that without strong prior information the simpler problem of deduction (using a model to predict data) could be substituted for the harder problem of induction (using data to predict a model) would determine the course of 20th-century science. The substitution is expressed formally as

$$\mathcal{L}(\theta|y) = p(y|\theta), \tag{7}$$

where $\mathcal{L}$ is used to represent the likelihood so that it is not confused with the posterior probability distribution $p(\theta|y)$. Absent any strong prior information, one can apply this substitution to infer the most likely model parameters θ given some data y. Because probability theory conforms with logic, the logical choice is to select, or at least prefer whatever model maximizes the likelihood function. This basic argument provides the basis for maximum likelihood estimation (MLE, Fisher, 1922), which are a class of methods for selecting the model θ having the greatest likelihood of having generated the data; formally,

$$\hat{\theta}_{MLE} = \text{argmax}_\theta(\mathcal{L}(\theta|y)), \tag{8}$$

where $\hat{\theta}_{MLE}$ represents the MLE estimate of θ. The justification of MLE leads directly to the justification of MAE and MSE, because under certain conditions the MAE and MSE are inversely proportional to the likelihood. That is to say, the model that minimizes the appropriate metric is also the most likely, but understanding exactly why this is so requires a bit more explanation.

3 The normal case

First, the case of normally distributed (Gaussian) errors. Consider a normally distributed variable y and some corresponding set of normally distributed model predictions $\hat{y}$. The model error is, therefore, the difference between two normal distributions. If y and $\hat{y}$ are independent, the error distribution is guaranteed to be ~~a~~ normal. Such a model provides no information, however, and for a model to be useful, $\hat{y}$ and y should be dependent. Although the difference between two dependent normal distributions is not guaranteed to be normal, it will often be so (Kale, 1970). Thus, we say that normally distributed variables will tend to produce normally distributed errors. As a starting point, assume the prediction errors are normal and independent and identically distributed (*iid*). ~~Subsequent sections will introduce ways~~ Ways of relaxing these assumptions are introduced in the next sections, but they provide a strong foundation, evident by the popularity of ordinary least squares. Our goal, then, is to identify the model $f()$ with normal *iid* errors that is most likely given the data, where $f()$ has inputs x and parameters θ, written as $f(x,\theta)$. The output of $f(x,\theta)$ is the model prediction $\hat{y}$, which represents the conditional mean of y given θ and x,

To find the most likely model, we begin with the likelihood given by the normal distribution,

$$\mathcal{L}(\mu,\sigma|y) = \frac{1}{\sqrt{2\pi\sigma^2}}\prod_{i=1}^{n}\exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right], \tag{9}$$

where μ is the population mean and σ is the standard deviation. Next, $f(\theta,x)$ is substituted for μ, replacing the population mean with the conditional mean,

$$\mathcal{L}(\theta,\sigma|y,x) = \frac{1}{\sqrt{2\pi\sigma^2}}\prod_{i=1}^{n}\exp\left[-\frac{(y_i-f(\theta,x_i)^2}{2\sigma^2}\frac{(y_i-f(\theta,x_i))^2}{2\sigma^2}\right]. \tag{10}$$

A ~~standard mathematical trick~~ convenient practice is to take the logarithm of the likelihood, thereby converting the products to sums

$$\log\mathcal{L} = -n\log\sigma - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-f(\theta,x_i))^2, \tag{11}$$

logging does not change the location of maximum, so it does not change the MLE estimate. From equation 11, it can be seen that maximizing the log-likelihood for the parameters θ is equivalent to minimizing the sum

$$\sum_{i=1}^{n}(y-f(\theta,x_i))^2, \tag{12}$$

which is the L2-norm. Dividing by n also has no effect the location of the maximum of the log-likelihood and yields the MSE. Thus, for normal *iid* errors, the model that minimizes the MSE (or the L2-norm) is the most likely model, all other things being equal. Although beyond our scope, information criteria, Bayesian methods, and cross validation are all techniques for dealing with situations where all other things are not equal and are closely related to topics discussed in this review.

4 The Laplace case

Now consider an exponentially distributed random variable; a concrete example being daily precipitation, which is often approximately exponential in distribution. If both model predictions and observations are *iid* exponential random variables,

then the model error will have a Laplace distribution (sometimes called a double exponential distribution). Like the normal case, such a model is not useful, so instead we focus on models for which predictions and observations are dependent. Such a model is not guaranteed to have Laplacian errors; nevertheless, its errors will tend to exhibit strong positive kurtosis, so we say it tends toward Laplacian-like error.

Assuming the Laplace distribution better represents the error than the normal, than we should prefer the model maximizing the Laplacian likelihood function,

$$\mathcal{L}(\theta, b|y, x) = \frac{1}{2b} \exp\left[-\frac{|y - f(\theta, x_i)|}{b}\right], \tag{13}$$

where b is a parameter of the distribution. Here we use the same substitution as in equation 10 to convert from the standard Laplace distribution to a Laplacian-error distribution. The log-likelihood is then

$$\log \mathcal{L} = -n \log(2b) - \frac{1}{b} \sum_{i=1}^{n} |y_i - f(\theta, x_i)|, \tag{14}$$

and repeating the argument from the normal case, maximizing the log-likelihood for θ is equivalent to minimizing the sum

$$\sum_{i=1}^{n} |y - f(\theta, x_i)|, \tag{15}$$

which is the L1-norm. Dividing the L1-norm by n yields the MAE. Thus, for Laplacian errors, the model that minimizes the MAE (or the L1-norm) also maximizes the likelihood.

5 Other options

To summarize the previous two sections: for normal errors, minimizing ~~(R)MSE~~ either MSE or RMSE yields the most likely model; whereas for Laplacian errors, minimizing MAE yields the most likely model. Normally distributed variables tend to produce normally distributed errors, and exponentially distributed variables tend to produce Laplacian-like errors, so RMSE and MAE are reasonable first choices for each case, respectively. Technically both also assume the errors are *iid*, and, for many interesting problems, errors are neither perfectly normal, nor Laplacian, nor *iid*. In these cases, there are essentially four options, all somewhat interrelated and often used in conjunction.

5.1 Refine the model structure

The first option is to refine the structure of the model; in other words, make the model more physically realistic. While this option is the most important for the advancement of science, it is not relevant to the choice of error metric, so I will not discuss it further, other than to note that ~~likelihood functions can~~ likelihoods can also be used to ~~inform incorporate decisions about model structure(Burnham and Anderson, 2001)~~ evaluate model structure: first, determine the maximum likelihood for each candidate model structure, then select (or prefer) the most likely among these candidates (Burnham and Anderson, 2001, e.g.,) . The preceding derivations were ~~all posed~~ formulated in terms of maximizing the likelihood by way of adjusting the model

parameters θ, but more generally, ~~θ can broadened to mean the entire model (structure and parameters)~~the likelihood can be used to refine the entire model—both its parameters and structure.

5.2 Transformation

The second option is to transform the data to a Laplace or, more commonly, normal distribution, then minimizing the MAE or RMSE of the transformed data will yield the most likely model. For example, the ~~distribution of streamflow at a particular location~~ streamflow distribution of a perennial stream is approximately lognormal. Logging a lognormal variable yields a normal ~~distribution, so~~ one, so in log space, the error is the difference between two normal distributions, which will also tend to be normal. If the errors can be made normal by transformation, then minimizing the MSE of the ~~log-transformed variable yields~~ transformed variable will yield the most likely model. Many statistical methods assume normality, and the general name for transforming a non-normal variable into a normal one is known as a Box-Cox transformation (Box and Cox, 1964). Transformations can make results harder to interpret, but this is usually an acceptable trade-off for better inference.

5.3 Robust inference

The third option is to use "robust" methods of inference. The term "robust" signifies that a technique is less sensitive to violations of its assumptions; typically, meaning they are less sensitive to extreme outliers. To achieve this, robust techniques replace the Gaussian likelihood with one with thicker-tails, such as the Laplace or the ~~Student-t~~Student's-t, which reintroduces the choice between RMSE and MAE, as MAE corresponds to the Laplace likelihood.

While Fisher (1920) demonstrated that minimizing the squared error was theoretically optimal for normal errors, he permitted Eddington to add a footnote that in practice, better results were often achieved by minimizing the absolute error, because in practice, observations include some outliers that ~~violate the normal condition~~ deviate from the normal distribution (Stigler, 1973). For this reason, minimizing the MAE has come to be known as a "robust" form of MLE, as in "robust regression" (e.g., Murphy, 2012, section 7.4). Tukey, in particular, was seminal in developing and exploring robust methods, such as in Tukey (1960) and his contributions to the field are documented by Huber (2002).

In their debate, Willmott et al. (2009) recognize robustness as an important advantage of MAE, though Chai and Draxler (2014) never directly acknowledge this point and instead advocate for "throwing out" outliers. Neither option is ideal. Either can yield reasonable results for minor deviations from the normal, but their performance degrades as the deviation grows.

Since Tukey's work, ~~a better alternative has emerged, known as~~ more robust alternatives have emerged, including the median absolute deviation or MAD,

$$\text{MAD} = b\,\text{median}_{i=1}^{n}|\hat{y}_i - \text{median}(y)|, \tag{16}$$

where typically $b = 1.483$ to reproduce the standard deviation in the case of the normal distribution (denoted as MAD_σ). Although MAD is less theoretically grounded, empirical evidence indicates it is more robust than MAE. MAD was first promoted by Hampel (1974), who attributed it to Gauss, and later by Huber (1981), and more recently by ~~?~~Gelman et al. (2020). One drawback is its relatively inefficiency for normal distributions (Rousseeuw and Croux, 1993), ~~though~~ but advocates of MAD

counter that RMSE is as ~~or more inefficient~~ inefficient (or more) for error distributions that deviate from the normal~~. Lacking a better alternative, ~~, so MAD remains a popular choice.

In addition to being "robust," MAD and MAE also preserve scale, unlike the formal likelihood-based approach discussed next. Unless combined with a transformation, scale-preserving error metrics have the same units as the data, such that their magnitude roughly corresponds to the magnitude of the typical error. While MAD and MAE are easy to interpret and implement, they are somewhat limited in scope in that they are only appropriate for "contaminated" distributions— mixtures of normal or Laplace distributions with a common midpoint, which, by implication, are also symmetric.

More complicated error distributions are frequently encountered in practice. For example, errors in rainfall-runoff models are typically heteroscedastic. Log ~~tranforming~~ transforming the data can correct this for positive streamflows values, but the log is undefined when streamflow is zero or negative. ~~Pragmatic~~ Simple work arounds, such as setting ~~zero~~ zeros to a small positive value, ~~can yield reasonable results~~ may be satisfactory when zero and near-zero values are relatively rare but blow up as those values become more frequent. Recall that in log space, errors are proportional, so the difference between 0.001 and 1 is ~~equivalent to the difference~~ the same as that between 1 and 1000.

5.4 Likelihood-based inference

The final option, likelihood-based inference, is the most versatile and subsumes the others in that each can be incorporated within its framework. Its main drawback is interpretative. The absolute value of the likelihood is meaningless, ~~neither does it measure of~~ unlike RMSE or MAE, which measure the typical error~~like RMSE or MAE~~. Their relative values are meaningful, however, in that ~~they represent the~~ the likelihood ratio represents the evidence for one model relative to another. ~~Log-likelihood is equivalent to the concept of entropy from information theory, so likelihood-based approaches are naturally combined with information-theory based ones.~~ For an accessible introduction to likelihood-based model selection, the reader is referred to Edwards (1992) and Burnham and Anderson (2001).

Metrics like RMSE and MAE are sometimes referred to as "informal" likelihoods because, in certain circumstances, they yield results equivalent to those obtained by the "formal" likelihood (e.g., Smith et al., 2008). Recall ~~from equation~~ that the model that minimizes the RMSE also maximizes the likelihood, if the errors are normal and *iid* (12). Informal likelihoods share some of the flexibility of formal ones, while preserving scale (commonly as real or percentage error). However, they have two notable drawbacks. Formal likelihoods are necessary when combining different distributions into one likelihood or when comparing among different error distributions ~~(e.g., normal versus Laplace Burnham and Anderson, 2001)~~ (e.g., normal versus Laplace; Burnham and Anderson, 2001). Furthermore, because informal likelihoods obscure their probabilistic origins, practitioners are frequently unaware of them and, as a consequence, use them incorrectly.

For streamflow modeling, examples of the formal likelihood-based approach include Schoups and Vrugt (2010); Smith et al. (2010, 2015). Schoups and Vrugt (2010) create a single flexible likelihood function with several parameters that can be adjusted to fit a range of complex error distributions; whereas, Smith et al. (2015) show the process of building complex likelihoods from combinations of simpler elements. Smith et al. (2015) focus on several variants of the zero-inflated normal distribution, which, in essence, inserts a normal likelihood within a binomial one. Additional components can be added to deal

with heteroscedasticity and serial dependence which are typical for errors in runoff modeling. For example, in the zero-inflated lognormal, a binomial component handles zeros values, while a log transformation handles heteroscedasticity in the positive values.

6 Why not use both RMSE and MAE?

Chai and Draxler (2014) argue for RMSE as the optimal metric for normal errors, refuting the idea that MAE should be used exclusively. They do not contend RMSE is inherently superior and instead advocate that a combination of metrics, including both RMSE and MAE, should be used to evaluate model performance. Many models are multi-faceted, so there is an inherent need for multi-faceted evaluation, but it can be problematic if approached ~~naively~~without considerable thought.

RMSE and MAE are not independent, so if both are presented, how should ~~the reader~~ we weigh their relative importance when evaluating a model? ~~One answer is to construct weights from their likelihood functions, the output of which represents the support or evidence for either metric (Burnham and Anderson, 2001).~~ Assuming no prior information, the logical approach is to weigh them 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, , p. 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 ~~unecessary~~ unnecessary and potentially confusing. ~~When evaluating multiple metrics, there is also a tendancy~~ If their evidence is similar, it may be appropriate to present a weighted average or present metrics along with their weights (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.

Although the likelihood can provide an objective measure of model performance, we are often concerned with multiple facets of a model, such that any one performance metric is insufficient. A common solution is to define and compute several metrics, each chosen to characterize a different aspect of the model's performance. For example, in rainfall runoff modeling a modeler may compute the error in flow volume (the model bias), as well the errors at a range of flow quantiles.

When evaluating these metrics, there is a tendency to combine them into an overall score, but such scores are not meaningful, at least in a maximum-likelihood sense. A better approach is to focus on ~~the "best" metric, then~~ a single objective function, like MSE for normally distributed errors, and decompose it into ~~independent~~ components representing specific aspects of a model's performance (e.g., Hodson et al., 2021).

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 equivalent to asking "how accurate is my model?", then comparing competing models, and selecting the most accurate. That first step—quantifying the uncertainty in a model— is import in its own right, especially if we base decisions on predictions from our models. Just as the Gaussian likelihood provides the theoretical basis for using RMSE to quantify model uncertainty when errors are normally distributed, more generally, other likelihood functions are used to evaluate model accuracy and confidence intervals for other error distributions.

7 Conclusions

Probability theory provides a logical answer to the choice between RMSE and MAE. Either metric is optimal in its correct application; though in practice, neither may be sufficient. For these cases, refining the model, transforming the data, using robust statistics, or constructing a better likelihood can yield better results. Arguably the latter is most versatile, though there are pragmatic reasons for preferring the others.

Returning to the debate over MAE and RMSE. Chai and Draxler (2014) were correct that RMSE is optimal for normally distributed errors, though ~~wrongly suggest~~they seem to suggest, wrongly, that MAE only applies to uniformly distributed errors. Whereas, Willmott and Matsuura (2005) and Willmott et al. (2009) were correct that MAE is more robust, though there are better alternatives. Most importantly, neither side provides the theoretical justification behind either metric, nor do they adequately introduce the extensive literature on this topic. Hopefully, this paper fills that gap by explaining why and when these metrics work, while ~~pointing to better~~ exposing readers to several alternatives for when they fail.

Code and data availability. No code or data were used in this manuscript.

Author contributions. TOH wrote the manuscript.

Competing interests. The author declares no competing intersects.

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

Acknowledgements. Funding for this research was provided by the Hydro-terrestrial Earth Systems Testbed (HyTest) project of the U.S. Geological Survey Integrated Water Prediction program.

References

Bernoulli, J.: Ars conjectandi: Usum & applicationem praecedentis doctrinae in civilibus, Moralibus & Oeconomicis, 4, 1713, 1713.

Boscovich, R. J.: De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impessa, Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii, 4, 353–396, 1757.

Box, G. E. and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society: Series B (Methodological), 26, 211–243, 1964.

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.

Clayton, A.: Bernoulli's Fallacy, Columbia University Press, https://doi.org/10.7312/clay19994, 2021.

Eddington, A. S.: Stellar movements and the structure of the universe, Macmillan and Company, limited, 1914.

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

Fisher, R. A.: A mathematical examination of the methods of determining the accuracy of observation by the mean error, and by the mean square error, Monthly Notices of the Royal Astronomical Society, 80, 758–770, https://doi.org/10.1093/mnras/80.8.758, 1920.

Fisher, R. A.: On the mathematical foundations of theoretical statistics, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222, 309–368, https://doi.org/10.1098/rsta.1922.0009, 1922.

Gauss, C. F.: Abhandlungen zur Methode der kleinsten Quadrate, chap. Bestimmung der Genauigkeit der Beobachtungen, pp. 185–196, 1816.

Gelman, A., Hill, J., and Vehtari, A.: Regression and other stories, Cambridge University Press, https://doi.org/10.1017/9781139161879, 2020.

Hampel, F. R.: The Influence Curve and its Role in Robust Estimation, Journal of the American Statistical Association, 69, 383–393, https://doi.org/10.1080/01621459.1974.10482962, 1974.

Hodson, T. O., Over, T. M., and Foks, S. F.: Mean squared error, deconstructed, Journal of Advances in Modeling Earth Systems, https://doi.org/10.1029/2021MS002681, 2021.

Huber, P. J.: Robust Statistics, John Wiley & Sons, Inc., https://doi.org/10.1002/0471725250, 1981.

Huber, P. J.: John W. Tukey's contributions to robust statistics, The Annals of Statistics, 30, https://doi.org/10.1214/aos/1043351251, 2002.

Jaynes, E. T.: Probability theory: The logic of science, Cambridge University Press, https://doi.org/10.1017/cbo9780511790423, 2003.

Kale, B. K.: Normality of linear combinations of non-normal random variables, The American Mathematical Monthly, 77, 992, https://doi.org/10.2307/2318121, 1970.

Laplace, P. S.: Théorie analytique des probabilités: Supplément a la théorie analytique des probabilités: Février 1818, Courcier, 1818.

Murphy, K. P.: Machine learning: a probabilistic perspective, MIT press, 2012.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. B.: Numerical recipes in C: The art of scientific computing, Cambridge University Press, 2 edn., 1992.

Rousseeuw, P. J. and Croux, C.: Alternatives to the Median Absolute Deviation, Journal of the American Statistical Association, 88, 1273–1283, https://doi.org/10.1080/01621459.1993.10476408, 1993.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resources Research, 46, https://doi.org/10.1029/2009wr008933, 2010.

Smith, P., Beven, K. J., and Tawn, J. A.: Informal likelihood measures in model assessment: Theoretic development and investigation, Advances in Water Resources, 31, 1087–1100, https://doi.org/10.1016/j.advwatres.2008.04.012, 2008.

Smith, T., Sharma, A., Marshall, L., Mehrotra, R., and Sisson, S.: Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, Water Resources Research, 46, https://doi.org/10.1029/2010wr009514, 2010.

Smith, T., Marshall, L., and Sharma, A.: Modeling residual hydrologic errors with Bayesian inference, Journal of Hydrology, 528, 29–37, https://doi.org/10.1016/j.jhydrol.2015.05.051, 2015.

Stigler, S. M.: Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher and the Discovery of the Concept of Sufficiency, Biometrika, 60, 439, https://doi.org/10.2307/2334992, 1973.

Stigler, S. M.: Studies in the History of Probability and Statistics. XL: Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation, Biometrika, 71, 615–620, https://doi.org/10.1093/biomet/71.3.615, 1984.

Tukey, J. W.: A survey of sampling from contaminated distributions, Contributions to probability and statistics, pp. 448–485, 1960.

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.

Willmott, C. J., Matsuura, K., and Robeson, S. M.: Ambiguities inherent in sums-of-squares-based error statistics, Atmospheric Environment, 43, 749–752, https://doi.org/10.1016/j.atmosenv.2008.10.005, 2009.