JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2021MS002681

Mean Squared Error, Deconstructed

Timothy O. Hodson[1], **Thomas M. Over**[1], **and Sydney S. Foks**[2]

[1]U.S. Geological Survey Central Midwest Water Science Center, USA, [2]U.S. Geological Survey Water Resources Mission Area, Tacoma, WA, USA

Key Points:

- Mean squared error (MSE) is an objective but somewhat enigmatic measure of model performance
- MSE can be decomposed into components that quantify specific aspects of model performance, such as bias and variance
- Mixing components among models yields a form of ensemble prediction

Correspondence to:

T. O. Hodson,
thodson@usgs.gov

Citation:

Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002681. https://doi.org/10.1029/2021MS002681

Received 15 JUL 2021
Accepted 21 NOV 2021

Author Contributions:

Conceptualization: Timothy O. Hodson, Thomas M. Over
Data curation: Timothy O. Hodson
Formal analysis: Timothy O. Hodson
Methodology: Timothy O. Hodson
Project Administration: Sydney S. Foks
Visualization: Timothy O. Hodson
Writing – original draft: Timothy O. Hodson
Writing – review & editing: Timothy O. Hodson, Thomas M. Over, Sydney S. Foks

Abstract As science becomes increasingly cross-disciplinary and scientific models become increasingly cross-coupled, standardized practices of model evaluation are more important than ever. For normally distributed data, mean squared error (MSE) is ideal as an objective measure of model performance, but it gives little insight into what aspects of model performance are "good" or "bad." This apparent weakness has led to a myriad of specialized error metrics, which are often aggregated to form a composite score. Such scores are inherently subjective, however, and while their components may be interpretable, the composite itself is not. We contend that, a better approach to model benchmarking and interpretation is to decompose MSE into interpretable components. To demonstrate the versatility of this approach, we outline some fundamental types of decomposition and apply them to predictions at 1,021 streamgages across the conterminous United States from three streamflow models. Through this demonstration, we hope to show that each component in a decomposition represents a distinct concept, like "season" or "variability," and that simple decompositions can be combined to represent more complex concepts, like "seasonal variability," creating an expressive language through which to interrogate models and data.

Plain Language Summary Models are essential scientific tools for explaining and predicting phenomena ranging from weather and climate, to health outcomes, to economic development, to the origins of the universe, and testing competing models is one of the most basic scientific activites. Yet, how scientists evaluate and justify their models can be inconsistent or even arbitrary. Traditionally, one performance metric—such as mean squared error—is used to identify the best model, but one metric provides little insight into what aspects of a model are "good" or "bad." This paper proposes a basic language for expressing different aspects of a model's performance. On one hand, this is useful for determining which aspects of model may require revision, but it also allows the modeler to separate out the best elements among several models and combine them to form an ensemble, analogous to how an audio engineer mixes together multiple tracks to form the best rendition of a musical piece.

1. Introduction

Whether used in physically based simulations or machine learning, error metrics provide essential benchmarks for model calibration (or training), verification, and validation. For models that predict a continuous variable, mean squared error (MSE) is an ideal performance benchmark because of its link to the concept of cross-entropy from information theory. Cross-entropy measures the similarity of two probability distributions. If the goal of modeling is to identify the model that most closely reproduces the true data-generating distribution, then the "best" model minimizes the cross-entropy between the model predictions and the training data (e.g., Kullback & Leibler, 1951). For normally distributed (Gaussian) data, minimizing the MSE is equivalent to minimizing the cross-entropy. Or in probabilistic terms, minimizing the MSE is equivalent to maximizing the likelihood of the data (Akaike, 1974, 1998; deLeeuw, 1992; Goodfellow et al., 2016). All other things being equal, Bayes' theorem proves that the model that maximizes the likelihood is also the most probable model given the data.

This link among MSE, cross-entropy, and likelihood applies to non-Gaussian data as well, so long as the data are transformed to a Gaussian distribution using a link function prior to calculating the MSE; two common link functions being the natural logarithm and logit. When used with a log link, the MSE becomes the mean squared logarithmic error (MSLE), which measures the relative difference between the true and predicted values. Another common manipulation is to normalize the MSE, most commonly by dividing it by the variance of the data. Normalization removes the effect of scale, allowing for comparison among models with multiple variables, like the Earth system models used for global climate prediction.

Through its link to cross-entropy, the MSE essentially compresses all the training data and model predictions into a single value quantifying how well a model reproduces reality. That strength has also been perceived as a weakness; by compressing so much information into a single number, the MSE provides little insight into what aspects of model performance are "good" or "bad." This has led to the development of multi-metric benchmarks, which are often combined as a composite score, as in the International Land Model Benchmarking Project (IL-AMB) system (Collier et al., 2018) and elsewhere (e.g., Efstratiadis & Koutsoyiannis, 2010; Gupta et al., 2009; Lindström et al., 1997). However, composite scores are inherently subjective, in that they depend on the choice of metrics and their weighting. Changing components results in an entirely new score, which is undesirable for a benchmark, and while their components may be more interpretable, the composite itself is not. A better approach is to decompose the MSE into interpretable components.

Decompositions have played several prominent roles among the earth sciences, including in the interpretation and composition of error metrics (e.g., Gupta et al., 2009; Murphy, 1988), as well as the early recognition by Searcy (1959) that streamflow records could be decomposed into volume and timing components. The latter is an example of the direct application of decomposition to describe data, as opposed to errors. Data decomposition can extract concepts from data for descriptive purposes or for predictive modeling. If one model better predicts streamflow timing and another better predicts volume, the best component from either model can be combined to generate a more accurate ensemble prediction (e.g., Farmer et al., 2018; Fennessey, 1994). Building upon these earlier domain-specific works, this paper outlines a general framework for separating, evaluating, and intermixing components of multiple data sets. Any framework needs a name; we refer to this one as "d-score," which is short for decomposed score.

2. Decomposition and Orthogonality

The model error, $\boldsymbol{\epsilon}$, is defined as,

$$\boldsymbol{\epsilon} = \hat{\boldsymbol{x}} - \boldsymbol{x} \tag{1}$$

where $\boldsymbol{x}$ is a vector of n observations and $\hat{\boldsymbol{x}}$ represents the corresponding model predictions. Like any quantity, error can be decomposed into components whose sum equals the total,

$$\boldsymbol{\epsilon} = \sum_{j=1}^{m} \boldsymbol{\epsilon}_j. \tag{2}$$

where $\boldsymbol{\epsilon}_j$ represents the jth of m error components. Just as errors can be decomposed, so can MSE. For two error components, represented by $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, the MSE decomposition is,

$$\mathrm{MSE}(\boldsymbol{\epsilon}) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \tag{3}$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\epsilon_{1i} + \epsilon_{2i})^2 \tag{4}$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\epsilon_{1i}^2 + \epsilon_{2i}^2 + 2\epsilon_{1i}\epsilon_{2i}) \tag{5}$$

$$= \mathrm{MSE}(\boldsymbol{\epsilon}_1) + \mathrm{MSE}(\boldsymbol{\epsilon}_2) + \frac{2}{n}\sum_{i=1}^{n} \epsilon_{1i}\epsilon_{2i} \tag{6}$$

where n is the number of observations, and the final term on the right-hand side of Equation 6 is analogous to the covariance between $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$. If $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are highly correlated, the sum of their products will tend to be large; whereas if they are uncorrelated, their products will tend to cancel when summed. Decomposition into m components is more algebraically complicated, but in practice, many useful decompositions are orthogonal, meaning their covariance is zero, so the decomposition simplifies such that the total MSE is the sum of the MSE of each component,

$$\mathrm{MSE}(\boldsymbol{\epsilon}) = \sum_{j=1}^{m} \mathrm{MSE}(\boldsymbol{\epsilon}_j). \tag{7}$$

Orthogonal decompositions are advantageous because the contribution from each component is separately quantifiable, and orthogonal decompositions can be combined to form compound decompositions that are also orthogonal, meaning that complex decompositions are easily formulated by combining simpler ones. A classic example being the decomposition of MSE into bias and variance (Geman et al., 1992),

$$\mathrm{MSE}(\boldsymbol{\epsilon}) = \mathbb{E}[\boldsymbol{\epsilon}^2] \tag{8}$$

$$= \left(\mathbb{E}[\boldsymbol{\epsilon}^2] - \mathbb{E}[\boldsymbol{\epsilon}]^2\right) + \mathbb{E}[\boldsymbol{\epsilon}]^2 \tag{9}$$

$$= \mathrm{Var}(\boldsymbol{\epsilon}) + \mathbb{E}[\boldsymbol{\epsilon}]^2 \tag{10}$$

$$= \mathrm{Var}(\boldsymbol{\epsilon}) + \mathrm{Bias}(\boldsymbol{\epsilon})^2, \tag{11}$$

where $\mathbb{E}[\boldsymbol{\epsilon}]$ is the mean of $\boldsymbol{\epsilon}$, called the bias, and $\mathrm{Var}(\boldsymbol{\epsilon})$ is its variance. The bias term quantifies how well the model reproduces the mean of the data, and the variance term quantifies how well the model reproduces variability in the data.

Although typically given for MSE, the same decomposition can be applied directly to any continuous random variable $\boldsymbol{x}$,

$$\boldsymbol{x} = (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]) + \mathbb{E}[\boldsymbol{x}] \tag{12}$$

where the third term on the right-hand side is the expected value of $\boldsymbol{x}$, and first two terms represent its variability. When applied to errors (substituting $\boldsymbol{\epsilon}$ for $\boldsymbol{x}$ in Equation 12), squaring these components and taking their mean yields back the MSE,

$$\mathrm{MSE}(\boldsymbol{\epsilon}) = \mathrm{Var}(\boldsymbol{\epsilon}) + \mathrm{Bias}(\boldsymbol{\epsilon})^2. \tag{13}$$

but more generally, the decomposition separates the concepts of mean and variability from any continuous random variable. In practical terms, this means that components can also be extracted from models or observations and then be recombined with components from other sources to form an ensemble prediction.

The Methods section outlines several more orthogonal decompositions; though brief, the selection is meant to highlight important motifs that will guide the reader in constructing others. A key insight is that each component in a decomposition represents a concept, like expected value, variance, trend, or seasonality, and that these general concepts can be combined to express more specific concepts, much like a language. When applied to errors, decompositions can be used to score how well a model reproduces particular concepts. Once the best components are identified among a suite of models, the same decompositions can be applied to separate those components from their models and combine them into an optimal ensemble prediction.

3. Methods

3.1. Computation of MSE

Streamflow is approximately lognormally distributed with heteroscedastic errors, so the data were log transformed prior to calculating the MSE in order to preserve the link between cross-entropy and MSE. Because streamflow may contain negative or zero values, which have no logarithm, streamflows less than 0.028 m^3/s (0.01 ft^3/s) were censored (observations and predictions less than 0.028 m^3/s were set to 0.028 m^3/s).

Note that in the computation of MSE (Equation 3) or its decompositions (e.g., Equation 6) n is the total number of observations, which may differ from the number of observations in a particular component. This is a subtle but important point. By setting n as the total number of observations, MSE ($\boldsymbol{\epsilon}_j$) represents the portion of the total MSE attributable to the error component $\boldsymbol{\epsilon}_j$. Consider a data set consisting of daily observations made over the course of one year. Decomposing the data by month would yield orthogonal components that each contain rough-

ly 30 daily observations. To calculate the contribution of each month to the total MSE, the n in Equation 3 should equal 365, not 30. In doing so we treat all error components as length n vectors, in which some elements may be zero. However, sometimes it is desirable to switch between conventions. When plotting model performance at multiple locations, we prefer to show the traditional MSE at each site (taking n as the number of observations at each site); otherwise, sites with fewer observations will appear to have smaller errors. However, when evaluating overall performance at multiple locations, n should equal the global n (the number of observations at all sites).

3.2. Scoring

Following the convention used by ILAMB, the MSE is transformed into a normalized score on the unit interval (Collier et al., 2018). First, the MSE is normalized, by dividing it by the variance of the observations; if a link function is used, divide by the standard deviation of the transformed data. Then, the score is computed by passing the normalized MSE (NMSE) through the exponential function,

$$\mathrm{NMSE}(\boldsymbol{\epsilon}) = \frac{\mathrm{MSE}(\boldsymbol{\epsilon})}{\sigma_x^2} \tag{14}$$

$$s = e^{-\alpha \mathrm{NMSE}(\epsilon)} \tag{15}$$

where σ_x is the standard deviation of the transformed observations, s is a score on the unit interval [0, 1] and α, set to 3.14 here, tunes the mapping of error to score (Collier et al., 2018). Scoring is purely esthetic in that it makes results more interpretable by humans without affecting how models rank relative to one another. The α parameter may be chosen to give scores an interpretable meaning (e.g., Collier et al., 2018) or, as done here, to distribute the scores across the unit interval, because well-distributed scores produce more informative plots.

Scoring the individual error components works the same as scoring the overall error. The MSE of orthogonal components sum to yield the total MSE. Similarly for scores, the product of the component score will equal the total score (because exponentiation converts sums to products). In other words, the total MSE and score are unaffected by the choice of decomposition, and components can be recombined to yield back the total. For easier visualization, we multiply the unit-interval scores by 100, so that 100 represents a perfect score. This scaling does not affect how models or their components rank relative to one another, though the component scores will no longer multiply to yield the total.

International Land Model Benchmarking Project takes a different approach to creating a total score, which is to average the component scores to create an overall score. As a result, the overall rankings of models depend on the choice of components, as well as on the choice of α, which is chosen somewhat arbitrarily. A more objective approach, which is proposed here, is to rank models based on the total MSE (or its scored form). The total MSE can then be decomposed into more informative components as necessary to evaluate specific aspects of model performance. For orthogonal components, their sum will equal the total MSE or, when scored, their product will equal the total score.

Although an ILAMB-style scoring function is used in this demonstration, decomposition works with other scoring functions as well, including the Nash-Sutcliffe efficiency (E) (Nash & Sutcliffe, 1970) or the coefficient of determination (R^2) defined as,

$$E = 1 - \mathrm{NMSE} \tag{16}$$

$$R^2 = 1 - \mathrm{NMSE}, \tag{17}$$

but one point of inconsistency arises in the formulation of E. When evaluating model performance at a specific location, E is a scored form of the MSE. The scoring is esthetic in that it does not affect the relative rankings of models, and the most plausible model always has the lowest MSE and, therefore, the highest E. However, when evaluating multiple locations, they are no longer equivalent, if, like in ILAMB, the NMSE at each location is calculated by dividing by the local variance (Collier et al., 2018). In order to preserve the relationship between score, MSE, and cross-entropy, the scores presented herein were normalized using the global variance, σ_x^2; that is, the variance of all the observations from all locations being evaluated.

3.3. Seasonal Decomposition

The simplest decompositions create components that are disjoint in time or space. If two components are disjoint, their errors will never coincide, so the components are orthogonal (their covariance is zero).

Decomposing errors by season is a basic example of a disjoint-in-time decomposition. The seasonal decomposition of error is,

$$\boldsymbol{\epsilon} = \sum_{j=1}^{m} \boldsymbol{\epsilon}_j \tag{18}$$

where j indicates the season: spring, summer, fall, or winter; and the seasonal components are given by,

$$\boldsymbol{\epsilon}_j = \boldsymbol{\delta}_j \boldsymbol{\epsilon} \tag{19}$$

$$\delta_{ij} = \begin{cases} 1 & \text{if } \epsilon_i \in \text{season } j \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where $\boldsymbol{\delta}_j$ is a length n indicator vector whose elements are equal to 1 if they correspond with season j or zero otherwise. The next section expands disjoint decompositions into the frequency domain.

3.4. Trend, Seasonality, and Residual Variability Decomposition

Another classical decomposition of time series data is into trend, seasonality, and residual variability,

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_s + \boldsymbol{\epsilon}_r \tag{21}$$

where m, s, and r denote the trend, seasonality, and residual variability, respectively. In general terms, this decomposition separates the data by frequency. Just as disjoint-in-time components are orthogonal, so too are disjoint-in-frequency components. Note the term "trend" can have other meanings, but in this context it is the sum of long-term variability and bias.

There are several methods of frequency decomposition. This paper uses STL, which stands for seasonal-trend decomposition using locally estimated scatterplot smoothing (Cleveland et al., 1990), with a 9-year seasonal smoothing window. Although the components produced by STL are not truly disjoint in frequency, they are close enough to be nearly orthogonal. STL is flexible in that it separates nonlinear trends and complex seasonal patterns, but that flexibility comes at a computational cost. Cheaper forms of frequency decomposition could use differencing to separate trends and the Yule-Walker equations to separate seasonality and variability, but these are also less flexible. Yet another approach is to Fourier transform the data (Cooley & Tukey, 1965), then perform the decomposition in the frequency domain. This is potentially fast and flexible but was unnecessary for this demonstration. Using approximately 100 cores, STL decomposition for the 1,021 streamflow error time series completed in a few minutes, but larger data sets may require a faster algorithm.

3.5. Bias, Distribution, and Sequence Decomposition

Recall that the variance component of the bias-variance decomposition represents how well the model reproduces variability in the data. Here, we show how the concept of variability can be decomposed further. Consider a hypothetical error time series formed by taking the difference between two sine waves. The bias component reflects differences in their means, and any remaining error reflects differences in their variability. For a sine wave, that remaining error must arise from some combination of differences in amplitude, frequency, or phase.

Although most models are not sine waves, the analogy offers an example of how variance can be decomposed into more specific concepts like frequency. The bias, distribution, and sequence decomposition provides an alternate decomposition of variance suitable for a broader variety of models. The basic idea is that the concept of variability can be decomposed into the sequence of events (timing) and their distribution (spread). Until this point,

decompositions have involved synchronous components, which means they compare model predictions against observations from the same point in time t,

$$\boldsymbol{\epsilon}_t = \hat{\boldsymbol{x}}_t - \boldsymbol{x}_t. \tag{22}$$

The bias, distribution, sequence decomposition introduces the more general case, in which the components may be asynchronous. The derivation begins by monotonically sorting the model predictions and observations, then decomposing the MSE of the result,

$$\boldsymbol{\omega} = \text{sort}(\hat{\boldsymbol{x}}) - \text{sort}(\boldsymbol{x}) \tag{23}$$

$$\text{MSE}(\boldsymbol{\omega}) = \text{Bias}(\boldsymbol{\omega})^2 + \text{Var}(\boldsymbol{\omega}). \tag{24}$$

Bias is invariant to sorting, so the bias term equals the bias of the unsorted errors, $\text{Bias}(\boldsymbol{\epsilon})$. As the sorted model predictions and observations share the same sequence, the variance between them, $\text{Var}(\boldsymbol{\omega})$, results from distributional error, $\text{Dist}(\boldsymbol{\epsilon})$,

$$\text{Var}(\boldsymbol{\omega}) = \text{Dist}(\boldsymbol{\epsilon}) \tag{25}$$

$$\text{MSE}(\boldsymbol{\omega}) = \text{Bias}(\boldsymbol{\epsilon})^2 + \text{Dist}(\boldsymbol{\epsilon}). \tag{26}$$

Furthermore, the only difference between $\text{MSE}(\boldsymbol{\epsilon})$ and $\text{MSE}(\boldsymbol{\omega})$ is that resulting from differences in their sequencing, $\text{Sequence}(\boldsymbol{\epsilon})$,

$$\text{MSE}(\boldsymbol{\epsilon}) - \text{MSE}(\boldsymbol{\omega}) = \text{Var}(\boldsymbol{\epsilon}) - \text{Var}(\boldsymbol{\omega}) \tag{27}$$

$$= \text{Sequence}(\boldsymbol{\epsilon}). \tag{28}$$

The full bias-distribution-phase decomposition is then,

$$\text{MSE}(\boldsymbol{\epsilon}) = \text{Bias}(\boldsymbol{\epsilon})^2 + \text{Var}(\boldsymbol{\epsilon}) \tag{29}$$

$$= \text{Bias}(\boldsymbol{\epsilon})^2 + (\text{Var}(\boldsymbol{\epsilon}) - \text{Var}(\boldsymbol{\omega})) + \text{Var}(\boldsymbol{\omega}) \tag{30}$$

$$= \text{Bias}(\boldsymbol{\epsilon})^2 + \text{Sequence}(\boldsymbol{\epsilon}) + \text{Dist}(\boldsymbol{\epsilon}). \tag{31}$$

A similar decomposition of MSE is given in Gupta et al. (2009) but with different sequence and distributional components.

3.6. Quantile Decomposition

In its basic form, quantile decomposition is another disjoint-in-time decomposition,

$$\boldsymbol{\epsilon} = \sum_{j=1}^{m} \boldsymbol{\epsilon}_j \tag{32}$$

where ϵ_j is the error component of the jth quantile given by,

$$\boldsymbol{\epsilon}_j = \boldsymbol{\delta}_j \boldsymbol{\epsilon} \tag{33}$$

$$\delta_{ij} = \begin{cases} 1 & \text{if } x_i \in \text{quantile } j \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

where $\boldsymbol{\delta}_j$ is an indicator vector equal to 1 where x_i falls within the jth quantile bin. Notice that the quantile bins are defined using the observations x_i; therefore, they evaluate how well different ranges in the magnitudes of the observations are predicted (at the time of their occurrence).

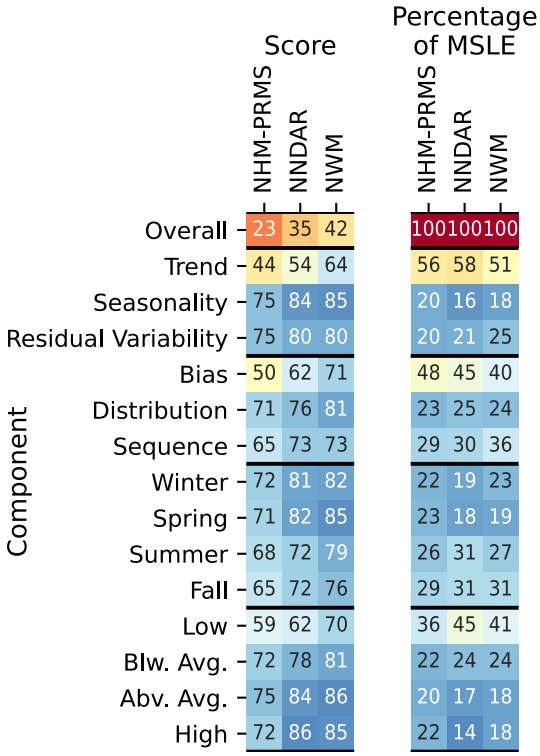

Figure 1. Scores (out of 100) and percentage of mean squared logarithmic error for each component of four decompositions. Reds and blues indicate worse and better performance, respectively. Low, below average (Blw. Avg.), above average (Abv. Avg.), and high components correspond to the first through fourth empirical quartiles, respectively. Black lines separate the four decompositions of the overall score.

Like any disjoint-in-time decomposition, the MSE can be decomposed into contributions from each quantile bin,

$$\mathrm{MSE}(\boldsymbol{\epsilon}) = \sum_{j=1}^{m} \mathrm{MSE}(\boldsymbol{\epsilon}_j). \tag{35}$$

For greater specificity, the quantile bins can be orthogonally decomposed further, and the resulting components will still sum to the MSE. We demonstrate this point by decomposing the quantile bins into their bias and variance,

$$\sum_{j=1}^{m} \mathrm{MSE}(\boldsymbol{\epsilon}_j) = \sum_{j=1}^{m} (\mathrm{Bias}(\boldsymbol{\epsilon}_j)^2 + \mathrm{Var}(\boldsymbol{\epsilon}_j)) \tag{36}$$

$$= \sum_{j=1}^{m} \mathbb{E}[\boldsymbol{\epsilon}_j]^2 + \sum_{j=1}^{m} \mathrm{Var}(\boldsymbol{\epsilon}_j). \tag{37}$$

The sum of the expectations is equal to the expectation of the sum, and for orthogonal components, the sum of the variances is equal to the variance of the sum, so the bias and variance of the quantile bins sum to reproduce the total MSE,

$$\sum_{j=1}^{m} \mathrm{MSE}(\boldsymbol{\epsilon}_j) = \mathbb{E}\left[\sum_{j=1}^{m} \boldsymbol{\epsilon}_j\right]^2 + \mathrm{Var}\left(\sum_{j=1}^{m} \boldsymbol{\epsilon}_j\right) \tag{38}$$

$$= \mathrm{Bias}(\boldsymbol{\epsilon})^2 + \mathrm{Var}(\boldsymbol{\epsilon}) \tag{39}$$

$$= \mathrm{MSE}(\boldsymbol{\epsilon}). \tag{40}$$

As noted, the quantile decomposition described in this section considers synchronous errors. In some applications, like the design of flood-control structures, the principal interest is how well a model reproduces the distribution of the data within a certain quantile bin, irrespective of how accurately it predicts the sequence (timing) of the data. These sorts of complex questions can be addressed using compound decompositions. By applying the quantile decomposition to the combined bias and distributional components, $\mathrm{Bias}(\boldsymbol{\epsilon}) + \mathrm{Dist}(\boldsymbol{\epsilon})$, the resulting quantile components quantify distributional errors that are independent of timing.

4. Data Set

As a demonstration, we applied error decomposition to three streamflow models: nearest-neighbor drainage area ratio (NNDAR), a simple statistical model that re-scales streamflow data from the nearest streamgage (e.g., Farmer et al., 2014); the version 3.0 calibration of the National Hydrologic Model Infrastructure application of the Precipitation-Runoff Modeling System (NHM-PRMS) (Hay & LaFontaine, 2020; LaFontaine et al., 2019); and version 2.0 of the National Water Model (NWM) (National Oceanic and Atmospheric Administration, 2016). The models were evaluated against streamflow observations from 1,021 "reference" (minimally anthropogenically impacted (Falcone, 2011)) watersheds across the conterminous United States with at least 10 years of observations.

5. Score Decomposition

Scores and percentage of the MSE applied to log-transformed streamflow (MSLE) for each error component are shown in Figure 1. The original unit-interval scores were multiplied by 100, so that 100 represents a perfect score. Because decomposition subdivides the total error into multiple components, each component contains less error, and therefore scores better, than the total. When left on the unit interval [0, 1], scores of orthogonal

Figure 2. Locations of the 1,021 "reference" streamgages in the conterminous United States. Streamgages at which nearest-neighbor drainage area ratio outperformed the National Water Model at predicting sequence are shown in black.

components multiply to equal the overall score. The trend, seasonality, and residual variability components are not quite orthogonal, so their product is less than the total.

Several factors bias these scores. Reference streamgages were not at randomly sampled locations, and many were used in calibrating the NHM-PRMS and NWM, such that their scores likely overestimate model performance at ungaged locations. Furthermore, reference streamgages tend to be clustered in space. As a result, NNDAR, which benefits from having nearby gages, likely scored higher than if the streamgages were distributed randomly. These limitations aside, the scores and their components are still useful for identifying strengths and weaknesses among the models.

The NWM had the highest overall score and outperformed the other two models in every component, except the highest quartile, where NNDAR was slightly better. There appears to be no benefit from borrowing components from the other two models to form an ensemble prediction, but other decompositions may yield more insight. For example, decomposing error by location shows that NNDAR generally outperforms NWM at predicting sequence in densely gaged regions (Figure 2). Therefore, an ensemble prediction might incorporate the sequence component of NNDAR just in those regions (i.e., order the NWM predictions in time according to the sequence of the NNDAR predictions (Fennessey, 1994)). However, the point of this example is not to prescribe a specific set of components, but rather to show the ease and flexibility with which new components are defined.

The overall score was calculated by averaging the MSLE of all sites, then passing the result through the scoring function (Methods). Using this approach, the worst performing sites tend to dominate in the overall score, which is expected because MSLE (and MSE, in general) is known for being sensitive to outliers. For example, the overall score of the NWM was 42, but the median site score was 79 (Figures 1 and 3). The effect of outliers could be reduced by scoring each site individually, then taking the average as the overall score (as in Collier et al., 2018). But in so doing, scoring becomes more than an esthetic operation; altering the scoring function (such as adjusting α) can alter how models rank relative to one another.

In traditional multi-metric benchmarking, a suite of metrics are used to evaluate various aspects of a model, but the choice of component metrics and their weighting are rather arbitrary. A model may score poorly on trend, even though trend matters little in the overall error. Seeing a low trend score, a modeler may focus on improving trend performance, which may do little for, or even be detrimental to, the overall performance. With decomposition, the error contribution from each component is clear (Figure 1). For the NWM, trend errors contribute more than half of the MSLE, more than twice that of seasonality or residual variability.

From the decompositions of NWM shown in Figure 1, the worst components of each decomposition were trend, bias, fall season, and low flows (51%, 40%, 31%, and 41% of MSLE, respectively). Being of different decompositions, these concepts are not necessarily orthogonal, so where should the modeler focus their attention? To better understand which matters more, we can decompose these concepts further to attempt to isolate orthogonal elements. In this case, the decomposition is trivial but instructive; the trend component—as defined by STL-includes bias, as well as long-term variability. Long-term variability can be isolated by taking the trend of the variance component or by taking the variance of the trend component. Decomposing trend in this manner shows that long-term variability contributed 11% of the MSLE, whereas bias contributed 40%, so bias had a larger effect on performance (to verify this, subtract the bias from the trend contribution to MSLE in Figure 1).

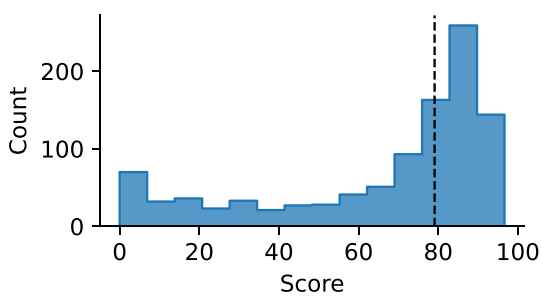

Figure 3. Overall scores for the National Water Model at 1,021 "reference" streamgages with at least 10 years of data. The dashed line indicates the median.

Scores can also be decomposed spatially and visualized as maps to provide additional clues about what physical processes may be contributing to model error (Figure 4). Although bias and variance make similar contributions to the MSLE of the NWM (40% and 60%, respectively), their spatial patterns are distinctly different. Bias errors were concentrated in the arid and

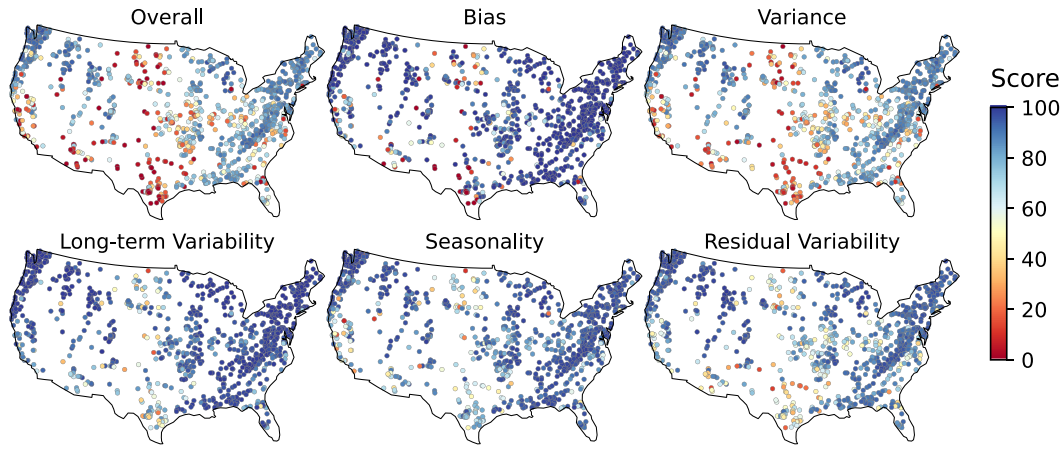

Figure 4. National Water Model scores at each streamgage decomposed as bias and variance, and variance further decomposed as long-term variability, seasonality, and residual variability.

semi-arid midcontinent (Figure 4). Variance errors were also large in these regions but were more widespread, specifically in the form of residual variability, which generally scored better in the mountainous regions of the eastern and western United States.

6. Conclusions

There is little-to-no comprehensive guidance for benchmarking model performance among scientific disciplines. Many disciplines—even individual researchers—develop their own ad hoc benchmarks, often with unclear meaning. Decomposition offers a better approach because it simultaneously allows flexibility in defining meaningful error metrics, while maintaining MSE as the principal benchmark of performance because of its link to cross-entropy and likelihood. Furthermore, by weighting the components and recombining them, decompositions can produce subjective benchmarks, useful for purpose-dependent calibrations. Or the same decompositions used in scoring can be applied to calibration data sets or model predictions to separate those data into components, which can be recombined with components from other data sets to create an ensemble prediction. Decomposition need not replace the myriad of discipline-specific error metrics, but it could serve as a common framework for evaluating models that increasingly bridge multiple disciplines.

At a basic level, modeling problems are categorized as either regression, that is, prediction of a quantity; and classification, that is, prediction of a label. We demonstrated several useful decompositions applicable to regression problems, which use MSE as a benchmark, but comparable decompositions exist for log-likelihood (Heskes, 1998), extending these concepts to classification problems. Just as simple words combine to express complex ideas, we envisage that simple decompositions can combine to represent complex concepts in any type of model or data. A universal language.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Data from NWM version 2.0 retrospective simulation are available at https://registry.opendata.aws/nwm-archive/. Data from NHM-PRMS are from Hay and LaFontaine (2020) and are available at https://doi.org/10.5066/P9PGZE0S. Data from NNDAR are from Russell et al. (2020) and are available at https://doi.org/10.5066/P9XT4WSP. Error component data are available in Hodson (2021) at https://doi.org/10.5066/P911RKX6.

Acknowledgments
This manuscript grew from constructive discussion among the members of the HyTest Evaluation Team, including Robert W. Dudley, Glenn A. Hodgkins, Julie E. Kiang, Sara B. Levin, Colin A. Penn, Amy M. Russell, Samuel W. Saxe, and Caelan E. Simeone. The authors also thank Gregory E. Schwarz and an anonymous reviewer for providing constructive feedback on this manuscript. 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. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike,* 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, *6*(1), 3–73.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., & Randerson, J. T. (2018). The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018ms001354

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, *19*(90), 297–301. https://doi.org/10.1090/s0025-5718-1965-0178586-1

deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. *Breakthroughs in Statistics*, 599–609. https://doi.org/10.1007/978-1-4612-0919-5_37

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal–Journal Des Sciences Hydrologiques*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Falcone, J. A. (2011). *Gages-II: Geospatial attributes of gages for evaluating streamflow*. U.S. Geological Survey dataset. https://doi.org/10.3133/ofr20111157

Farmer, W. H., Archfield, S. A., Over, T. M., Hay, L. E., LaFontaine, J. H., & Kiang, J. E. (2014). *A comparison of methods to predict historical daily streamflow time series in the southeastern United States*. U.S. Geological Survey Scientific Investigations Report. https://doi.org/10.3133/sir20145231

Farmer, W. H., Over, T. M., & Kiang, J. E. (2018). Bias correction of simulated historical daily streamflow at ungauged locations by using independently estimated flow duration curves. *Hydrology and Earth System Sciences*, *22*(11), 5741–5758. https://doi.org/10.5194/hess-22-5741-2018

Fennessey, N. M. (1994). *A hydro-climatological model of daily streamflow for the northeast United States* (Unpublished doctoral dissertation). Tufts University.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from http://www.deeplearningbook.org

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hay, L., & LaFontaine, J. (2020). *Application of the National Hydrologic Model infrastructure with the precipitation-runoff modeling system (NHM-PRMS), 1980-2016, Daymet version 3 calibration*. U.S. Geological Survey data release. https://doi.org/10.5066/P9PGZE0S

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, *10*(6), 1425–1433. https://doi.org/10.1162/089976698300017232

Hodson, T. (2021). *Mean squared logarithmic error in daily mean streamflow predictions at gages-ii reference streamgages*. U.S. Geological Survey data release. https://doi.org/10.5066/P911RKX6

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

LaFontaine, J. H., Hart, R. M., Hay, L. E., Farmer, W. H., Bock, A. R., Viger, R. J., & Driscoll, J. M. (2019). *Simulation of water availability in the Southeastern United States for historical and potential future climate and land-cover conditions*. U.S. Geological Survey Scientific Investigations Report. https://doi.org/10.3133/sir20195039

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1–4), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM<2.0.CO;2

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

National Oceanic and Atmospheric Administration. (2016). *The National water model*. Retrieved from https://water.noaa.gov/about/nwm

Russell, A. M., Over, T. M., & Farmer, W. H. (2020). *Cross-validation results for five statistical methods of daily streamflow estimation at 1,385 reference streamgages in the conterminous United States, water years 1981-2017*. U.S. Geological Survey Data Release. https://doi.org/10.5066/P9XT4WSP

Searcy, J. K. (1959). *Flow-duration curves* (No. 1542-A). U.S. Geological Survey Water Supply Paper. https://doi.org/10.3133/wsp1542a