Performance Metrics, Error Modeling, and Uncertainty Quantification

YUDONG TIAN
Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, and Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

GREY S. NEARING
Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, and Science Applications International Corporation, Beltsville, Maryland

CHRISTA D. PETERS-LIDARD
Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

KENNETH W. HARRISON
Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, and Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

LING TANG
Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

(Manuscript received 13 March 2015, in final form 18 November 2015)

ABSTRACT

A common set of statistical metrics has been used to summarize the performance of models or measurements—the most widely used ones being bias, mean square error, and linear correlation coefficient. They assume linear, additive, Gaussian errors, and they are interdependent, incomplete, and incapable of directly quantifying uncertainty. The authors demonstrate that these metrics can be directly derived from the parameters of the simple linear error model. Since a correct error model captures the full error information, it is argued that the specification of a parametric error model should be an alternative to the metrics-based approach. The error-modeling methodology is applicable to both linear and nonlinear errors, while the metrics are only meaningful for linear errors. In addition, the error model expresses the error structure more naturally, and directly quantifies uncertainty. This argument is further explained by highlighting the intrinsic connections between the performance metrics, the error model, and the joint distribution between the data and the reference.

1. Limitations of performance metrics

One of the primary objectives of measurement or model evaluation is to quantify the uncertainty in the data. This is because the uncertainty directly determines the information content of the data (e.g., Jaynes 2003), and dictates our rational use of the information, be it for data assimilation, hypothesis testing, or decision-making. Further, by appropriately quantifying the uncertainty, one gains insight into the error characteristics of the measurement or model, especially via efforts to separate the systematic error and random error (e.g., Barnston and Thomas 1983; Ebert and McBride 2000).

Currently the common practice of measurement or model verification is to compute a common set of performance metrics. These performance metrics are statistical measures to summarize the similarity and difference between two datasets. These metrics are based on direct comparison of datum pairs on their corresponding spatial/temporal location. The most commonly used ones are bias, mean square error (MSE), and correlation coefficient (CC) (e.g., Fisher 1958; Wilks 2011), but many
variants or derivatives, such as (un)conditional bias (e.g., Stewart 1990), MSE (Murphy and Winkler 1987), unbiased root-mean-square error (ubRMSE), anomaly correlation coefficient, coefficient of determination (CoD), and skill score (SS; e.g., Murphy and Epstein 1989) also fall into this category. Table 1 lists some of these metrics and their definitions.

Among them, the “big three”—bias, MSE, and CC—are the most widely used in diverse disciplines, exemplified by the popular “Taylor diagram” (Taylor 2001).

These metrics do, however, have several limitations:

1) Interdependence. Most of these conventional performance metrics are not independent; they have been demonstrated to relate to each other in complex ways. For example, the MSE can be decomposed in many ways to link it with other metrics, such as bias and correlation coefficient (e.g., Murphy 1988; Barnston 1992; Taylor 2001; Gupta et al. 2009; Entekhabi et al. 2010). These relations indicate both redundancy among these metrics, and the metrics’ indirect connection to independent error characteristics. This leads to ambiguity in the interpretation and intercomparison of these metrics.

2) Underdetermination. It is easy to verify that these metrics do not describe unique error characteristics, even when many of them are used collectively. In fact, many different combinations of error characteristics can produce the same values of these metrics.

This is illustrated in Fig. 1. A monthly time series of land surface temperature anomaly data, extracted from satellite-based observations (Wan et al. 2004) over a location in the United States (35°N, 95°W), is used as the reference (black curves) for validating two separate hypothetical sets of predictions (Figs. 1a and 1b, blue curves). Their respective scatterplots are also shown (Figs. 1c and 1d), with values of five major conventional metrics listed (bias, MSE, CC, CoD, and SS). When seen from either the time series plots or the scatterplots, the two measurements exhibit apparently very different error characteristics. However, all the metrics, except bias, give nearly identical values (Figs. 1c and 1d). In fact, there is an infinite number of ways to construct measurements that can produce identical values for many of the metrics. Therefore, when given a set of these metrics values, one will have fundamental difficulty in inferring and communicating the error characteristics of the predictions.

3) Incompleteness. There are no well-accepted guidelines on how many of these metrics are sufficient. Many inexperienced users follow a “the more the better” philosophy, and it is not rare to see works

---

**Table 1. Examples of conventional performance metrics.** The observations and forecasts are denoted as $x$ and $y$, respectively.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Ideal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i)$</td>
<td>0</td>
</tr>
<tr>
<td>Mean square error (MSE)</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i)^2$</td>
<td>0</td>
</tr>
<tr>
<td>Linear correlation coefficient (CC)</td>
<td>$\frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}}$</td>
<td>1</td>
</tr>
<tr>
<td>Coef of determination (CoD)</td>
<td>$\frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$</td>
<td>1</td>
</tr>
<tr>
<td>MSE skill score (SS) (also known as Nash–Sutcliffe efficiency in hydrology)</td>
<td>$1 - \frac{\sum_{i=1}^{N} (y_i - x_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$</td>
<td>1</td>
</tr>
</tbody>
</table>

*Notations used to denote sample mean and variance for $x$ and $y$, respectively, are defined as

\[
x_i \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} x_j, \quad y_i \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} y_j, \quad \sigma^2_x \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} (x_j - \bar{x})^2, \quad \text{and} \quad \sigma^2_y \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} (y_j - \bar{y})^2.
\]
that employ as many as 7–10 metrics, many being variants of one another. However, end users and stakeholders generally prefer as concise a summary as possible.

4) Linear error assumption. The metrics discussed here are all based on the least squares measure. The underlying assumption is that the measurement model is linear, with additive, Gaussian random error. When this assumption is not satisfied, one can still compute these metrics, yet they are misleading or meaningless (Barnston and Thomas 1983; Habib et al. 2001). In fact, these metrics have been frequently abused when one applies them in ignorance of the underlying error model.

Many enhanced techniques and metrics have been proposed to overcome these limitations. These include various decomposition schemes of the conventional metrics (e.g., Murphy 1988; Stewart 1990; Barnston 1992; Taylor 2001; Gupta et al. 2009; Entekhabi et al. 2010), object-oriented metrics (e.g., Ebert and McBride 2000), and nonparametric metrics that do not rely on underlying assumptions about the nature of the systematic error (e.g., Weijs et al. 2010; Gong et al. 2013; Vrugt and Sadegh 2013). These are all generally valuable efforts; however, we doubt that these kinds of relatively complex statistical analyses will ever fully replace the use of linear performance metrics. As such, we propose an alternative philosophy when applying these simple metrics.

2. Error characterization with an error model

If we denote the reference observations and predictions by \( x \) and \( y \), respectively, then the most generic mathematical expression for the relationship between

---

**FIG. 1.** Dramatically different measurement errors can produce identical values of the conventional metrics. (a),(b) Two separate measurements (blue curves) of the same reference dataset (black curves) are constructed, and (c),(d) their associated scatterplots are shown (blue line: linear fit of measurements vs truth; black line: linear fit of truth vs truth). The corresponding values of the conventional metrics (bias, MSE, CC, CoD, and SS) are given in each scatterplot. The two erroneous measurements, distinctly different, yield nearly identical values (subject to errors due to limited sample size) in all the metrics but bias. In contrast, the three parameters \((\lambda, \delta, \sigma)\) of the error model in (1) clearly depict their differences.
the two is their joint distribution, \( p(y, x) \). As Murphy et al. (1989) pointed out, \( p(y, x) \) “contains all of the non-time-dependent information relevant to forecast verification.” Therefore, the most general approach for model verification is directly characterizing \( p(y, x) \) (Nearing and Gupta 2015). Since

\[
p(y, x) = p(y | x)p(x),
\]

and \( p(x) \) is given, this is equivalent to characterizing \( p(y | x) \), which is the conditional distribution, or the likelihood function, or the error model. Murphy and Winkler (1987) named this equation “likelihood-base rate factorization.” We hereby argue and demonstrate that the error model and the conventional metrics are closely related, and that there are unique advantages offered by the error model.

The simplest error model is the linear, additive error model. It is the most widely used or assumed (sometimes incorrectly):

\[
y_i = \lambda x_i + \delta + \epsilon_i,
\]

where \( x_i \) and \( y_i \) are the observations and predictions, respectively, indexed by \( i = 1, 2, \ldots, N \), with \( N \) being the total number of data points. The three terms on the right-hand side represent three separate sources of estimation errors. The first term, with an error parameter \( \lambda \), represents scale error, or dynamic-range error, when \( \lambda \) deviates from unit. The second term, \( \delta \), represents a constant deviation, and can be called the displacement error. The third term, \( \epsilon_i \), is a stochastic variable and indicates the random error, typically assumed as independent and identically distributed, with zero mean value and a standard deviation of \( \sigma \).

The first two terms depict a deterministic relationship between \( x \) and \( y \), and they jointly quantify the systematic error, with \( \lambda \) and \( \delta \). The random error, \( \epsilon_i \), a stochastic term, is solely quantified by \( \sigma \), and is “orthogonal” (Papoulis 1984) to the systematic error. The separation of the errors into systematic and random parts is critical for uncertainty quantification, as random error is an amalgamation of all types of errors unexplainable with current data and knowledge at hand, and directly reflects our current state of ignorance.

It is easy to see that (1) characterizes the conditional distribution \( p(y | x) \) as \( p(y | x) \sim N(\lambda x + \delta, \sigma^2) \), where \( N(\cdot) \) denotes the Gaussian distribution, assumed here for convenience. Therefore, an error model is essentially a parametric description of the conditional distribution. Subsequently, if (1) is the correct error model, its three parameters, \( \lambda \), \( \delta \), and \( \sigma \), capture all the error characteristics in the measurements. Given any set of estimates \( y_i \) and reference \( x_i \), the three quantities, \( \lambda \), \( \delta \), and \( \sigma \), and their confidence intervals, can be easily estimated in any number of ways, including the ordinary least squares method, the maximum likelihood method, or the Bayesian method (e.g., Carroll et al. 2006; Wilks 2011).

Once the error model (1) is parameterized, it is easy to see that all the relevant error information is captured by the three parameters. This can be demonstrated in Fig. 1 as well. While for the two sets of measurements shown in Figs. 1a and 1b, four of the five conventional metrics fail to differentiate them (see their respective values in Figs. 1c and 1d), the three model parameters completely captured the different error characteristics, including their differences in the scale error \( (\lambda = 0.6 \text{ vs } 1.7) \), displacement error \( (\delta = 6 \text{ vs } 5.1) \), and random error \( (\sigma = 1.1 \text{ vs } 3.3) \). Moreover, their representation of the error characteristics is strikingly intuitive, when one compares the numbers with the scatterplots (Figs. 1c and 1d). For example, the scale error \( \lambda \), being less than 1 for the first set of measurements, indicates the measurements compress the dynamic range of the data, a fact easily confirmed by both the time series (Fig. 1a) and the scatterplot (Fig. 1c). The same effect applies to the second measurements. The popular conventional metrics listed in Table 1, in contrast, clearly lack such explanatory power.

As mentioned above, the conventional linear performance metrics can be computed directly from our three parameters. Table 2 lists several such derivations. For example, the conventional bias can be written as

\[
\text{bias} = \overline{y} - \overline{x} = (\lambda - 1)\overline{x} + \delta.
\]

Bias contains both the contribution from the scale error (first term on rhs) and the displacement error (second term). Even with no displacement error present \( (\delta = 0) \), the mere presence of scale error \( (\lambda \neq 1) \) will be interpreted as “bias.” Further, bias reflects the combination of both inherent flaws in the model \( (\lambda \text{ and } \delta) \) and the mean value of the data sample \( \overline{x} \). Therefore, a thermometer will have a different bias for measuring daytime or nighttime temperatures, for example. This is misleading, and may affect the development of “bias correction” strategies. In contrast, \( \lambda \) and \( \delta \) define the error character of the predictions (in the context of the linear error model), independent of any particular data samples; once computed with a given calibration dataset, they will—assuming validity of the error model—remain valid in the context of other observations.

The MSE can be expressed as

\[
\text{MSE} = (\overline{y} - \overline{x})^2 + (\lambda - 1)^2\sigma_\epsilon^2 + \sigma^2,
\]

which indicates MSE has contributions from three components: the mean error (ME), the scale error, and the random error. Note that the mean error is actually composed of the scale error of the reference mean and the
displacement error, as shown in (2). Therefore, MSE is a mix of all the three types of errors—the main cause for its lack of independence and explanatory power. Knowledge of MSE alone, therefore, does not offer guidance on how the predictions or measurements might be improved. It is interesting to compare (3) with the various conventional MSE decomposition schemes. One of them is (Murphy 1988; Gupta et al. 2009; Entekhabi et al. 2010; Wilks 2011):

$$\text{MSE} = (\bar{y} - \bar{x})^2 + \sigma_x^2 + \sigma_y^2 - 2\sigma_x \sigma_y \text{CC}. \quad (4a)$$

Murphy (1995) proposed a variation:

$$\text{MSE} = (\bar{y} - \bar{x})^2 + (\sigma_y - \sigma_x \text{CC})^2 + (1 - \text{CC}^2)\sigma_x^2. \quad (4b)$$

Willmott (1981) suggested yet another scheme:

$$\text{MSE} = \delta^2 + (\lambda - 1)^2 \frac{1}{N} \sum_{i=1}^{N} x_i^2 + 2\delta(\lambda - 1)\bar{x} + \sigma^2, \quad (4c)$$

and he names the first two terms additive and proportional systematic errors, the third term is the interdependence between them, and the last one is the random error. These decomposition schemes definitely help solve the underdetermination issue shown in Fig. 1, and deserve wider adoption. It is also notable that the components in any of the schemes above do not enjoy complete mutual independence. All these schemes are mathematically identical, but the formulation here in (3) is apparently the simplest and most intuitive, demonstrating the benefit of connecting the conventional schemes with the underlying error modeling.

The linear CC can also be expressed with the model parameters. It is the square root of a ratio (Table 2). The ratio is “signal” over “signal + noise.” Or if we define a signal-to-noise ratio (SNR) as

$$\text{SNR} = \frac{\lambda \sigma_y}{\sigma}, \quad (5)$$

then it is easy to see

$$\text{CC}^2 = \frac{1}{1 + 1/\text{SNR}^2}, \quad (6)$$

which means when \( \text{SNR} \rightarrow 0, \text{CC}^2 \rightarrow 0, \) and when \( \text{SNR} \rightarrow \infty, \text{CC}^2 \rightarrow 1. \) This is quite intuitive, and again the information CC produces is already contained in the underlying error model via \( \lambda \) and \( \sigma. \)

As far as all the performance metrics illustrated in the Taylor diagram are concerned, the three parameters of the error model are sufficient to fully capture the error characteristics of model predictions. There is no additional error information to be gained from these performance metrics.

In fact, the relationship between the conventional metrics and the linear, additive error model (1) has been noted in earlier works (e.g., Murphy 1995; Livezey et al. 1995). However, they invariably failed to explicitly define and treat the random error with a parametric form. This is the root cause of the complexity in the interpretation of the conventional metrics, and the vagueness in using them to quantify uncertainty. The various MSE decomposition schemes [(4a)–(4c)] help alleviate the problem, yet a decomposition scheme with the error-parameter representation (3) is more straightforward.

Of course the linear error model (1), albeit widely adopted, is not always appropriate. Some situations demand nonlinear error models. For example, for high-resolution

---

**Table 2. Expressions of the conventional performance metrics with the three model parameters.***

<table>
<thead>
<tr>
<th>Name</th>
<th>Derivation from the three new metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>( (\lambda - 1)\bar{x} + \delta )</td>
</tr>
<tr>
<td>Mean square error (MSE)</td>
<td>( [(\lambda - 1)\bar{x} + \delta]^2 + (\lambda - 1)^2\sigma_x^2 + \sigma^2 )</td>
</tr>
<tr>
<td>Linear correlation coefficient (CC)</td>
<td>( \lambda \frac{\sigma_y}{\lambda \sigma_x + \sigma} )</td>
</tr>
<tr>
<td>Coef of determination (CoD)</td>
<td>( \frac{\lambda^2 \sigma_y}{\lambda^2 \sigma_x^2 + \sigma^2} )</td>
</tr>
<tr>
<td>MSE skill score (SS) (also known as Nash–Sutcliffe efficiency in hydrology)</td>
<td>( 1 - \frac{[(\lambda - 1)\bar{x} + \delta]^2 + (\lambda - 1)^2\sigma_x^2 + \sigma^2}{\sigma_x^2} )</td>
</tr>
</tbody>
</table>

* Notations used to denote sample mean and variance for \( x \) and \( y \), respectively, are defined as

\[
\bar{x} \equiv \frac{1}{N} \sum_{i=1}^{N} x_i, \quad \bar{y} \equiv \frac{1}{N} \sum_{i=1}^{N} y_i, \quad \sigma_x^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2, \quad \text{and} \quad \sigma_y^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2.
\]
precipitation estimates, a nonlinear, multiplicative error model proves to be more appropriate (Tian et al. 2013; Tang et al. 2015):

\[ y_i = \alpha x_i^\beta \varepsilon_i. \]  \hspace{1cm} (7)

Nevertheless, once the analytical form of the error model is given, the error information is fully captured by the model parameters. In the case of the nonlinear error model (7), the systematic error is represented by a nonlinear function of \( \alpha \) and \( \beta \), and the random error, \( \varepsilon_i \), is no longer additive and Gaussian. For the particular case of precipitation errors studied in (Tian et al. 2013), the probability distribution of the random error is often assumed approximately lognormal:

\[ \varepsilon_i \sim \text{LN}(0, \sigma^2), \]

and the parameter \( \sigma \) quantifies the random error.

However, for such a nonlinear, non-Gaussian error model, or any generic nonlinear error models, the conventional metrics are inappropriate; they are not meaningful representations of the error structure (Barnston and Thomas 1983) as their assumed separation of systematic and random error no longer applies. For example, Habib et al. (2001) demonstrated that the correlation coefficient would produce misleading results in such cases. Therefore, the indiscriminate use of the conventional metrics with nonlinear error models such as (7) is imprudent and incorrect.

Error characterization is equivalent to characterizing the conditional distribution of model predictions given observations (the error model), while the conventional metrics are only valid for the linear, additive error model (1). Therefore, we argue that one should base their error analysis on the appropriate error model, which defines and separates the systematic and random errors, and its parameters quantify them. It should be up to the researcher to demonstrate that a particular error model fits well for this purpose. Of course, the choice of the correct error model is dictated by the model developer’s empirical understanding of the error characteristics.

3. Summary

Model verification is a statistical inference process. During the verification period when both the observation \( x \) and model forecast \( y \) are available, one will be able to establish the conditional distribution, or error model, \( p(y \mid x) \). This will enable one to rationally estimate the observation \( x \) and the associated uncertainty when only forecast \( y \) is available, \( p(x \mid y) \), such as in many real-time applications (e.g., Tian et al. 2010). This step is enabled by Bayes’s theorem (e.g., Jaynes 2003):

\[ p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}. \]  \hspace{1cm} (8)

Thus, model verification is built upon the correct characterization of the error model \( p(y \mid x) \). Of course, this inference procedure between \( p(y \mid x) \) and \( p(x \mid y) \) is fully reversible [see (8)], depending on the focus of the evaluation (Murphy and Winkler 1987).

Based on this consideration, here we promote a philosophy for error analysis and uncertainty quantification that puts more emphasis on error modeling than on the conventional performance metrics computing. There are two primary motivations for this comment.

First, the most popular conventional performance metrics: bias, MSE, and correlation coefficients, have several limitations related to interdependence, non-determination, and incompleteness. These limitations make it difficult to infer the error characteristics with a given set of metrics values (Fig. 1), and their roles in characterizing \( p(y \mid x) \) are very vague. The traditional approach to alleviate this problem is to introduce more metrics or decomposition schemes [e.g., (4a)–(4c)], leading to a proliferation of metrics. We demonstrate that these metrics and decomposition schemes can be, however, all derived from the linear error model, which can be interpreted natively in a way that is both coherent and free from these difficulties, without invoking a multitude of metrics.

Second, making the first step in the process of model evaluation the choice of error model, forces us to reconcile our specific knowledge of the situation, including research and application goals, with the way we choose to present the reconciliation of our predictive models with observational data.

It is worth emphasizing that the validity of both error modeling and performance metrics, or the entire verification process as a whole [see (8)], hinges on the assumption of statistical stationarity. Thus, when we establish an error model or compute a set of metrics, we should select the appropriate spatial and temporal domains within each of which the error characteristics of the instrument or model are believed to be stationary, and treat each domain (e.g., summer vs winter) with separate model fitting or metrics computing.

It is important to be cognizant of the fact that the error model itself is a characterization of a conditional probability distribution \( p(y \mid x) \), which contains all of the information available for evaluating our predictive model (Murphy et al. 1989). Our job is to find a suitable analytic, parametric description of the conditional. Since
the error information is distilled in the error model’s parameters, it is not surprising to see the relationship between most of the conventional metrics and the model parameters (Table 2). Of course, a suitable parametric error model is not always easy to find, and in such cases, nonparametric approaches, which are fundamentally based on information theoretic concepts (Nearing and Gupta 2015; Weijs et al. 2010) may be employed.

Acknowledgments. This research was supported by the NASA Earth System Data Records Uncertainty Analysis Program (Martha E. Maiden) under solicitation NNH10ZDA001N-ESDRERR. Computing resources were provided by the NASA Center for Climate Simulation. We thank two anonymous reviewers for their helpful comments.

REFERENCES