# Assessment of the Adequacy of Mathematical Models

Luis Orlindo Tedeschi[1]
Department of Animal Science
Cornell University, Ithaca, NY 14850

## 1. Abstract

Models are mathematical representations of mechanisms that govern natural phenomena that are not fully recognized, controlled, or understood. They are an abstraction of the reality. Models have become indispensable tools via decision support systems for policy makers and researchers to provide ways to express the scientific knowledge. Model usefulness has to be assessed through its sustainability for a particular purpose. Nonetheless, model testing is often designed to demonstrate the rightness of a model and the tests are typically presented as evidences to promote its acceptance and usability. Adequate statistical analysis is an indispensable step during all phases of model development and evaluation because a model without statistics is like a body without a soul. Therefore, in this paper we discussed and compared several techniques for mathematical model comparison. It was concluded the identification and acceptance of wrongness of a model is an important step towards the development of more reliable and accurate models. The assessment of the adequacy of mathematical models is only possible through the combination of several statistical analyses and proper investigation regarding the purposes for which the mathematical model was initially conceptualized and developed for. The use of few techniques may be misleading in selecting the appropriate model in a given scenario.
**Keywords**: Modeling, Simulation, Evaluation, Validation, Adequacy

## 2. Introduction

Models are mathematical representations of mechanisms that govern natural phenomena that are not fully recognized, controlled, or understood. Mathematical modeling has become an indispensable tool via computerized decision support systems for policy makers and researchers to provide ways to express the scientific knowledge, to lead to new discoveries, and/or to challenge old dogmas.

The notion that all models are wrong but some are useful (Box, 1979) is bothersome at first sight. However, the understanding and acceptance of the wrongness and weaknesses of a model strengthens the modeling process, making it more resilient and powerful in all aspects during the development, evaluation, and revision phases. Rather than ignoring the fact that a model may fail, one should incorporate the failures of a model in the learning process and exploit the principles in developing an improved redesigned model. In systems thinking, the understanding that models are wrong and humility about the limitations of our knowledge is essential in creating an environment in which we can learn about the complexity of systems in which we are embedded (Sterman, 2002).

Model testing is often designed to demonstrate the rightness of a model and the tests are typically presented as evidences to promote its acceptance and usability (Sterman, 2002). Models can be evaluated rather than validated in the strict sense of establishing truthfulness if one accepts the concept that all models are wrong. One should focus on the modeling process instead of model results per se. The modeling process encompasses several steps that begin with a clear statement of the objectives of the model, assumptions about the model boundaries, appropriateness of the available data, design of the model structure, evaluation of the simulations, and providing feedback for recommendations and redesign processes.

[1] Email: lot1@cornell.edu

Two procedures for model development are commonly used: (a.) develop conceptual ideas and interactions, and then perform the parameterization of the variables of the model with suitable data, or (b.) analyze experimental data that explain a biological phenomenon, and then combine them in a methodological manner. The former will generate a conceptual (or theoretical) model whereas the second one will be an empirical model.

When developing a model, one should make the best decision despite the foreseeable confines of our scientific knowledge and current modeling techniques. Adequate statistical analysis is an indispensable step during all phases of model development simply because a model without (proper) statistics is like a body without a soul.

A computer program containing the statistical procedures for model evaluation presented in this review may be downloaded at http://www.cncps.cornell.edu/modeval.

## 3. Evaluation of Mathematical Models

The evaluation of model adequacy is an essential step of the modeling process because it indicates the level of precision and accuracy of the model predictions. This is an important phase either to build up confidence on the current model or to allow selection of alternative models. Forrester (1961) emphasizes that the validity of a mathematical model has to be judged by its sustainability for a particular purpose; that means, it is a valid and sound model if it accomplishes what is expected of it. Shaeffer (1980) developed a methodological approach to evaluate models that consisted of six tasks: (a.) model examination, (b.) algorithm examination, (c.) data evaluation, (d.) sensitivity analysis, (e.) validation studies, and (f.) code comparison studies. A comprehensive list of publications regarding model *validation* was compiled by Hamilton (1991).

### *3.1. Conflicts in the Modeling Terminology*

Terminology use and its interpretation are often discussed when considering the appropriateness of mathematical models. The concept of *validation* or *verification* has been strongly criticized because it is philosophically impossible to prove that all components of models of real systems are true or correct (Oreskes et al., 1994).

Since models are an abstraction and representation of the reality, they can never fully mimic the reality under all conditions; and in fact, some models may be programmed to predict quantities or relationships that cannot be measured or observed in the real system; therefore, they cannot be *validated*.

Harrison (1991) differentiates *verification* from *validation* as follows; *verification* is designed to ensure that a mathematical model performs as intended, while *validation* examines the broader question of whether the intended structure is appropriate. Similarly, Sterman (2000) indicates that *validation* means "having a conclusion correctly derived from premises" while *verification* implies "establishment of the truth, accuracy, or reality of". Nonetheless, regardless of the definition adopted, none of these terms can support modeling *validation* or *verification*.

Hamilton (1991) proposes that *validation* is used to assess the extent to which a model is rational and fulfills its purposes. It is comprised of three tasks: (1.) verification (design, programming, and checking processes of the computer program), (2.) sensitivity analysis (behavior of each component of the model), and (3.) evaluation (comparison of model outcomes with real data).

There is the misuse of the word *calibration*. The concept of *calibration* means model fine tuning or fitting; it is the estimation of values of parameters or unmeasured variables using (proper) available information from the real system.

Models can substantiate an assumption or hypothesis by offering evidence that would strengthen concepts that may be already fully or partly established through other means (Oreskes et al., 1994) such as experimentation or observational studies. Therefore, the terms *evaluation* or *testing* are proposed to indicate measurement of model adequacy (or robustness) based on pre-established criteria

of model performance acceptance such as functionality, accuracy, and precision for its intended purpose.

Regardless of the objective of a model, a functional model is an abstraction of the reality and an approximation of the real system. Therefore, the evaluation phase is used to determine whether the model is an adequate representation for the process it was designed for rather than establishing the truth of the model in any absolute sense. Mathematical models cannot be proven valid, only whether it is appropriate for its intended purpose for given conditions.

### *3.2. Errors in the Modeling Process*

As shown in Table 1, the acceptance of a model is subjected to a selection process that may lead to an erroneous decision. The type I error (rejecting an appropriate model) is likely to occur when biased or incorrect observations are chosen to evaluate the model. For instance, these observations may result from the lack of adequate adjustment for random factors or covariate. A type II error (accepting an inappropriate model) happens when biased or incorrect observations were used during the model development phase, and assuming that independent incorrect observations were used to evaluate the model; therefore, the model will reflect it. The occurrence of a type II error can be as bad as or worse than the type I error.

Both types of errors can be minimized with appropriate statistical analysis to detect model appropriateness and the use of unbiased observations during both the development and evaluation phases. Therefore, failure to prove a significant difference between observed and model-predicted values may only be due to insufficient replication or lack of power of the applied statistical test (Mayer and Butler, 1993) rather than the correctness of a particular model.

Table 1. Two-way decision process in model evaluation

| | Model Predictions | |
|---|---|---|
| Decision | Correct | Wrong |
| Reject | Type I Error ($\alpha$) | Correct ($1 - \beta$) |
| Accept | Correct ($1 - \alpha$) | Type II Error ($\beta$) |

The thorough evaluation of a mathematical model is a continuous and relentless process. The fact that one or a series of evaluation tests indicated the mathematical model is performing adequately implies nothing about the future predictions of the mathematical model regarding its prediction ability. On the contrary, as the redesign (or revision) of the mathematical model progresses, the evaluation phase becomes more crucial.

### *3.3. The Concepts of Accuracy and Precision*

*Accuracy* measures how closely model-predicted values are to the true values. *Precision* measures how closely individual model-predicted values are within each other. In other words, accuracy is the model's ability to predict the right values and precision is the ability of the model to predict similar values consistently. Figure 1 illustrates the difference between accuracy and precision using the analogy of a target practice.

*Inaccuracy* or *bias* is the systematic deviation from the truth; therefore, even though the points in case 1 (Figure 1) are more sparsely distributed than in case 2 (Figure 1), both are equally inaccurate (or biased) because they are not centered in the target bull's eye. In contrast, *imprecision* or *uncertainty* refers to the magnitude of the scatter about the average mean; therefore, even though the points in case 3 and 4 (Figure 1) are equally accurate, case 3 has a higher imprecision than case 4 because the points in case 4 are more tightly grouped. Similarly, Figure 2 depicts the concepts illustrated in Figure 1 in a numerical form.

In Figure 2, the X-axis and Y-axis represent model-predicted and observed values, respectively. The regression estimate of coefficient of determination ($r^2$) is a good indicator of precision: the higher the $r^2$ the higher the precision. The regression estimates of the intercept and the slope are good indicators of accuracy; the simultaneously closer to zero and unity, respectively, the higher the accuracy.

Therefore, precision and accuracy are two independent measures of prediction error of a model: high precision does not guarantee high accuracy. The question then becomes *which measure is better?* It could be argued that precision is not as important as accuracy because the true mean can be detected using an imprecise method simply by averaging a large number of data points (case 3 in Figure 1). On the other hand, it could be argued that precision is more important than accuracy because the true mean is irrelevant when detecting differences among model predictions. It is easy to understand why accuracy could be argued as the most important measure because it measures the ability of a model to predict true values. The judgment of model adequacy always has to be made on relative comparisons and it is subject to the suitability of the model given its objectives. The lack of accuracy shown in case 2 (Figure 2) is also known as translation discrepancy (Gauch et al., 2003); despite the slope being equal to unity, the intercept is different from zero.

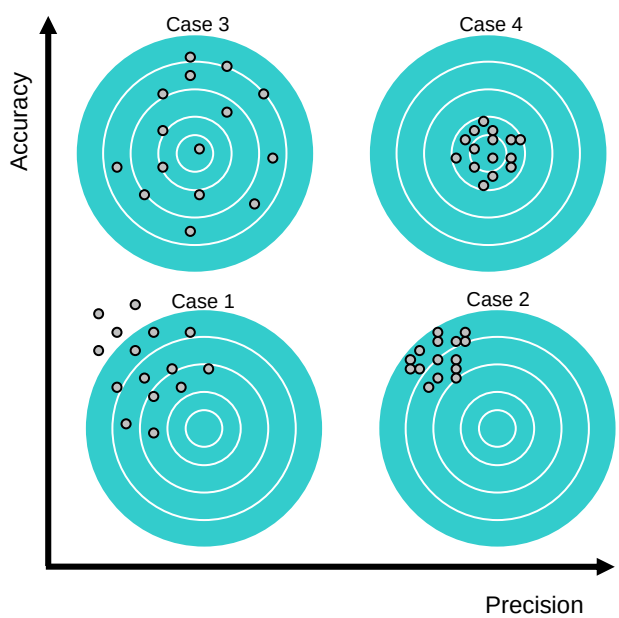
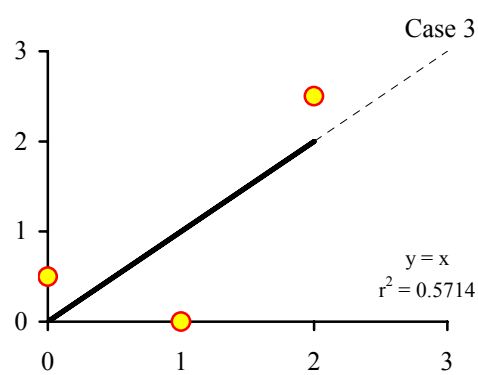
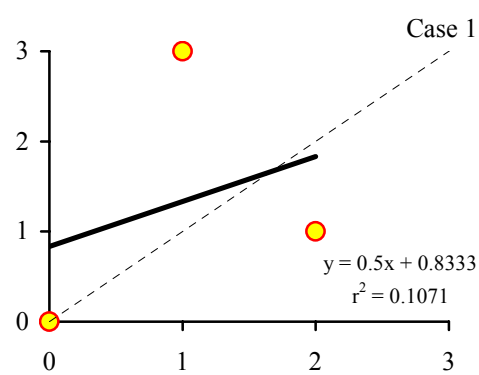



Figure 1. Schematization of accuracy versus precision. Case 1 is inaccurate and imprecise, case 2 is inaccurate and precise, case 3 is accurate and imprecise, and case 4 is accurate and precise.



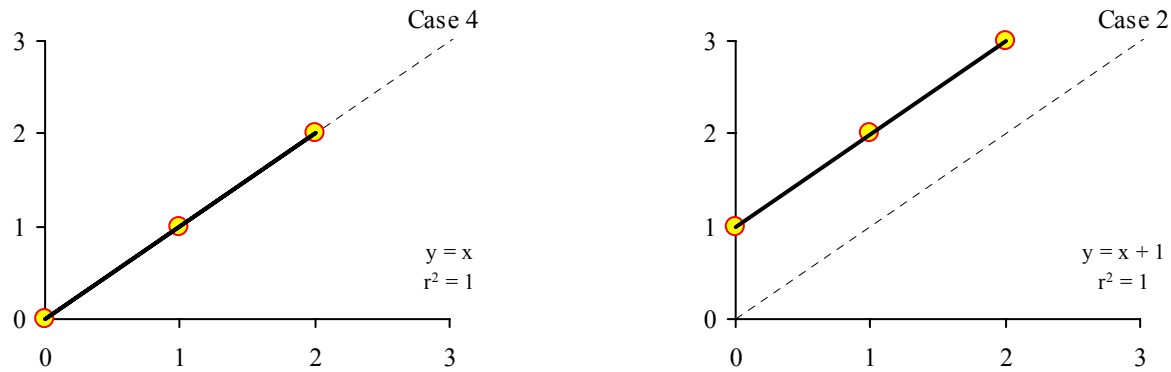



Figure 2. Comparison of accuracy and precision measures. Case 1 is inaccurate and imprecise, case 2 is inaccurate and precise, case 3 is accurate and imprecise, and case 4 is accurate and precise. The dotted line represents the Y = X line.

## 4. Techniques for Model Evaluation

### *4.1. Analysis of Linear Regression*

Table 2 shows the mathematical notations used throughout this paper. Note that *X* is the input variable used in the mathematical model.

The model-predicted values ( $f(X_1,...,X_p)_i$ ) are plotted in the X-axis while the observed values are plotted in the Y-axis because the observed values contain natural variability whereas the model-predicted values are deterministic with no random variation (Harrison, 1990; Mayer and Butler, 1993; Mayer et al., 1994). In this plot format, data points lying below and above the Y = X line indicate over- and under-prediction by the mathematical model, respectively. In contrast, Analla (1998) states that both variables ($Y_i$ and $f(X_1,...,X_p)_i$) are random; hence, it does not matter which variable is to be regressed on the other. Stochastic models can be re-run many times, decreasing the error of the mean value, which in practice does not invalidate the linear regression technique (Gunst and Mason, 1980). However, as discussed above, unless the mathematical model is considered stochastic (or probabilistic), the model-predicted values are fixed, deterministic, and associated with no variance; consequently, $Y_i$ should be regressed on $f(X_1,...,X_p)_i$.

Table 2. Mathematical notations, acronyms, and variable descriptions

| Notations | Description |
|---|---|
| $Y_i$ | This is the i$^{th}$ observed or measured value |
| $\bar{Y}$ | Mean of the observed (or measured) values |
| $\hat{Y}$ | Predicted values by the statistical regression model |
| $f(X_1,...,X_p)_i$ | This is the i$^{th}$ model-predicted (or simulated) value |
| $\bar{f}(X_1,...,X_p)$ | Mean of model-predicted (or simulated) values |
| $X_1,...,X_p$ | Input variables for the mathematical model |
| Deviation | It is the difference between model-predicted and observed values ($Y_i - f(X_1,...,X_p)_i$) |
| Residue (or error) | It is the difference between observed and regression-predicted values ($Y_i - \hat{Y}_i$) |

Several quantitative approaches involving statistical analysis have been used to evaluate model adequacy. A linear regression between observed and predicted values is commonly used. The hypothesis is that the regression passes through the origin and has a slope of unity (Dent and Blackie, 1979). Nonetheless, the use of the least-square method to derive a linear regression of observed on model-predicted values for model evaluation has little interest since the predicted value is useless in evaluating the mathematical model; therefore, the $r^2$ is irrelevant since one does not intend to make predictions from the fitted line (Mitchell, 1997).

Additionally, necessary assumptions have to be considered when performing a linear regression: (a.) the X-axis values are known without errors; this is true if and only if the model is deterministic and the model-predicted values are used in the X-axis, (b.) the Y-axis values have to be independent, random, and homoscedasticity, and (c.) residuals are independent and identically distributed $\sim \mathbb{N}\left(0,\sigma^2\right)$. Shapiro-Wilk's W test is frequently used to ensure normal distribution of the data (Shapiro and Wilk, 1965). Equation [1] has the general format of a first-order equation representing a linear regression used for model evaluation.

$$Y_i = \beta_0 + \beta_1 \times f(X_1,...,X_p)_i + \varepsilon_i \qquad [1]$$

Where $Y_i$ is the i$^{th}$ observed value, $f(X_1,...,X_p)_i$ is the i$^{th}$ model-predicted value, $\beta_0$ and $\beta_1$ are the regression parameters for the intercept and the slope, respectively; and $\varepsilon_i$ is the i$^{th}$ random error (or residue) that is independent and identically distributed $\sim \mathbb{N}\left(0,\sigma^2\right)$.

The estimators of the regression parameters $\beta_0$ and $\beta_1$ (Equation [1]) are computed using the least-square method (Neter et al., 1996) in which the square of the difference between expected or regression-predicted value ($\hat{Y}$) and the observed value ($Y$) is minimized, as shown in Equation [2]. Therefore, the estimators of regression parameters $\beta_0$ and $\beta_1$ are those values $a$ and $b$ that minimize $Q$ (Equation [2]).

$$Q = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 \times f(X_1,...,X_p)_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad [2]$$

The values of $a$ and $b$ can be algebraically computed by a sequence of calculations. The mean for observed and model-predicted values are computed as shown in Equation [3] and [4], respectively.

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad [3]$$

$$\overline{f}(X_1,...,X_p) = \frac{\sum_{i=1}^{n} f(X_1,...,X_p)_i}{n} \qquad [4]$$

The slope ($b$) and the intercept ($a$) of the linear regression are computed as shown in Equation [5] and [6], respectively.

$$b = \frac{\sum_{i=1}^{n}\left[\left(Y_i - \overline{Y}\right)\times\left(f(X_1,...,X_p)_i - \overline{f}(X_1,...,X_p)\right)\right]}{\sum_{i=1}^{n}\left(f(X_1,...,X_p)_i - \overline{f}(X_1,...,X_p)\right)^2} \qquad [5]$$

$$a = \overline{Y} - b \times \overline{f}(X_1,...,X_p) \qquad [6]$$

The sample standard deviation for $Y$ and $f(X_1,...,X_p)$ are computed as indicated in Equations [7] and [8], respectively.

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n-1}} \qquad [7]$$

$$s_{f(X_1,\ldots,X_p)} = \sqrt{\frac{\sum_{i=1}^{n}\left(f(X_1,\ldots,X_p)_i - \overline{f}(X_1,\ldots,X_p)\right)^2}{n-1}} \qquad [8]$$

The (co)variance is computed as indicated in Equation [9].

$$s_{f(X_1,\ldots,X_p)Y} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(Y_i - \overline{Y}\right)\times\left(f(X_1,\ldots,X_p)_i - \overline{f}(X_1,\ldots,X_p)\right)\right] \qquad [9]$$

The Person's coefficient of correlation ($r$) is computed as indicated in Equation [10]. The coefficient of determination ($r^2$) is computed as $r \times r$ using Equation [10].

$$r = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{s_{f(X_1,\ldots,X_p)Y}}{s_Y \times s_{f(X_1,\ldots,X_p)}} \qquad [10]$$

The $r$ coefficient, and its counterpart $r^2$, requires careful interpretation. There are several misunderstandings surrounding this statistic as discussed by Neter et al. (1996): (a.) a high coefficient of correlation does not indicate that useful predictions can be made by a given mathematical model since it measures precision and not accuracy, (b.) a high coefficient of correlation does not imply that the estimated regression line is a good fit of the model prediction since the relation can be curvilinear, and (c.) an $r$ near zero does not indicate that observed and model-predicted values are not correlated because they may have a curvilinear relationship.

The unbiased estimator of the variance of the random error ($\sigma^2$) is the mean square error, or residual mean square, or standard error of the estimate (MSE or $s_{y.x}$), which is computed using Equation [11]. Analla (1998) proposes the proof of model validity should be based on the MSE when comparing several mathematical models.

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} \qquad [11]$$

Note that MSE can be computed from the estimates of $r$ (Equation [10]) and $s_Y^2$ (Equation [7]) with adjustments for degrees of freedom as shown in Equation [12].

$$MSE = \frac{s_Y^2 \times (n-1) \times \left(1 - r^2\right)}{n-2} \qquad [12]$$

The standard deviations of the intercept and slope estimates can be computed as shown in Equations [13] and [14], respectively.

$$s_a = \sqrt{MSE \times \left(\frac{1}{n} + \frac{\overline{f}(X_1,\ldots,X_p)^2}{\sum_{i=1}^{n}\left(f(X_1,\ldots,X_p)_i - \overline{f}(X_1,\ldots,X_p)\right)^2}\right)} \qquad [13]$$

$$s_b = \sqrt{\frac{MSE}{\sum_{i=1}^{n}\left(f(X_1,...,X_p)_i - \overline{f}(X_1,...,X_p)\right)^2}} \quad [14]$$

The studentized statistics of the intercept and slope is *t*-distributed with ($n$ – 2) degrees of freedom. Therefore, the confidence interval to check for intercept ≠ 0 or slope ≠ 1 can be computed as: $a \pm t_{(1-\alpha/2,n-2)} \times s_a$ and $(b-1) \pm t_{(1-\alpha/2,n-2)} \times s_b$, respectively.

Unfortunately, the *t*-test for the slope has an ambiguous interpretation because the more scatter in the data points, the greater is the standard error of the slope, the smaller is the computed value for the statistical test and therefore, the harder it is to reject the null hypotheses, which states the slope is equal to unity. Therefore, the test can fail to reject the null hypothesis either because the slope is really not different from unity or there is much scatter around the line (Harrison, 1990). Alternatively, the confidence interval should be used to investigate the range of the slope (Mitchell, 1997).

Dent and Blackie (1979) have suggested a more relevant null hypothesis that tests if the intercept and the slope coefficients simultaneously are not different from zero and unity, respectively. Equation [15] shows the appropriate statistic calculation using the *F*-test with 2 and ($n$ – 2) degrees of freedom. Mayer et al. (1994) showed the derivation of Equation [16], which is similar to Equation [15] without the term ($n$ – 2) in the numerator and the term ($n$) in the denominator. This adjustment may cause different acceptance threshold Equations [15] and [16] based on the probability of the *F*-test.

$$F_1 = \frac{(n-2)\times\left(n\times a^2 + 2\times n\times \overline{f}(X_1,...,X_p)\times a\times(b-1) + \sum_{i=1}^{n}\left(f(X_1,...,X_p)_i^2 \times (b-1)^2\right)\right)}{2\times n\times MSE} \quad [15]$$

$$F_2 = \frac{n\times a^2 + 2\times n\times \overline{f}(X_1,...,X_p)\times a\times(b-1) + \sum_{i=1}^{n}\left(f(X_1,...,X_p)_i^2 \times (b-1)^2\right)}{2\times MSE} \quad [16]$$

Where $n$ is the sample size, $a$ is the intercept of the linear regression, $b$ is slope of the linear regression, $\overline{f}(X_1,...,X_p)$ is the mean of model-predicted values, and $f(X_1,...,X_p)_i$ is the i$^{th}$ model-predicted value.

The *F*-test calculation presented in Equation [15] is valid only for deterministic models; in fact the slope should not be expected to be unity when using stochastic (or probabilistic) models. Nonetheless, as shown by Analla (1998) the greater the MSE (Equation [11]) the more difficult it is to reject $H_0$ that intercept and slope simultaneously differ from zero and unity, respectively; a larger confidence interval is expected, which increases its tolerance.

Additionally, the independently distributed assumption may cause some problems since many evaluation datasets are time series or autocorrelated in some other way (Harrison, 1990), theoretically resulting in the expected values of the slope and the intercept of the linear regression being less than one and greater than zero, respectively. Therefore, the simultaneous *F*-test may not perform acceptably (Mayer et al., 1994). The remedial for autocorrelation is the average of subsequent pairs of data (or triplets) within the time series; unfortunately, there is loss of power of the *F*-test as the number of observations decreases (Mayer et al., 1994).

Because of disagreements of wrongness and rightness of the regression of observed on model-predicted values, Kleijnen et al. (1998) proposed a sequence of analysis consisting of the regression of the difference between model-predicted and observed values on their associated sums to evaluate model adequacy. Stringent assumptions require that a model is valid if and only if the observed and the model-predicted values have (a.) identical means, (b.) identical variances, and (c.) they are positively correlated.

There are several concerns about the appropriateness of linear regression in model evaluation (Harrison, 1990; Mitchell, 1997). The assumptions are rarely satisfied, the null hypothesis tests give ambiguous results depending on the scatter of the data, and regression lacks sensitivity because distinguishing lines between random clouds or data points is not necessary at this stage in model development. Regression analysis tests the slope equal to unity, which can reject a reasonably good agreement if the residual errors are small.

#### *4.1.1. Uncertainties in the Y- and X-Variates*

When the mathematical model is stochastic (or probabilistic) or when an inherent variability or uncertainty of the predicted values is assumed, the least-square method is not appropriate for estimating the parameters of the linear regression. Mayer et al. (1994) indicated that if the X-variate contains some error, the rejection rates of valid models may increase up to 27% to 47% depending on whether the errors are uncorrelated or correlated, respectively.

A solution to this inadequacy is the use of Monte Carlo techniques to provide goodness-of-fit tests regardless of within-model peculiarities (Shaeffer, 1980; Waller et al., 2003). The Monte Carlo test verifies if the observed data appear consistent with the model in contrast to whether the model appears consistent with the observed data.

Another solution is the use of least square regression weighted by the standard deviation of the data points. The $\chi^2$ merit function is represented in Equation [17] (Press et al., 1992). Unfortunately, an iterative process is required to minimize Equation [17] with respect to *a* and *b* since the *b* parameters occur in the denominator, which makes the resulting equation for the slope $\partial\chi^2/\partial b = 0$ nonlinear. Lybanon (1984) has provided an algorithm to minimize Equation [17]; additional comments to the procedure have also been provided (Jefferys, 1980, 1981, 1988a, b; Reed, 1989, 1990, 1992; Squire, 1990).

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(Y_i - a - b \times f(X_{1,\ldots,} X_p)_i\right)^2}{\sigma_{Y_i}^2 + b^2 \times \sigma_{f(X_{1,\ldots,} X_p)_i}^2} \qquad [17]$$

Where $\sigma_{Y_i}^2$ and $\sigma_{f(X_{1,\ldots,} X_p)_i}^2$ are the variance of the i$^{th}$ data point for observed and model-predicted values, respectively.

### *4.2. Analysis of Fitting Errors*

#### *4.2.1. Analysis of Deviations*

This is an empirical evaluation method proposed by Mitchell (1997) and Mitchell and Sheehy (1997). In this case, the deviations (model-predicted minus observed values) are plotted against the observed values (or model-predicted values as in Mitchell, 1997) and the percentage of points lying within an acceptable range (envelope) is used as a criterion of model adequacy. The acceptable range is based on the purpose of the model; usually the limits of 95% confidence interval of the observed values are used as a guideline. Ultimately, the differences between observed and model-predicted values provide adequate information on how far the model fails to simulate the system (Mitchell and Sheehy, 1997).

#### *4.2.2. Analysis of Residuals: Extreme Points*

Analysis of residuals (regression-predicted minus observed values)) is an important process in identifying data points that causes departures of the assumptions considered in the linear regression. Such data points are known as outliers or extreme and the study of their behavior is exceptionally important in assessing the aptness of the model and/or the measures necessary to improve the adequacy

of the model. Therefore, analysis of residuals also has an important role in model development and in the redesign phases.

The diagnostics for residues have been comprehensively discussed by Neter et al. (1996, Ch. 3 and 9). The departures of nonlinearity of the regression and nonconstancy of error variance is assessed by the plot of residual versus the predictor variable or the plot of residual versus the fitted values.

*Leverage value*. The detection of outliers or extreme data points is not a simple or an easy task. Nonetheless, it is very important in studying the outlying cases carefully and deciding whether they should be retained or removed from the dataset. There are several effective statistics that assist in the detection of these misbehavior data points. Most of these statistics use the leverage value ($h_{ii}$), which is computed from the diagonal element of the hat matrix. Equation [18] shows the special case of the calculation of the leverage values for the linear regression.

$$h_{ii} = \left(c_1 + c_2 \times f(X_1,...,X_p)_i\right) + \left(c_2 + c_3 \times f(X_1,...,X_p)_i\right) \times f(X_1,...,X_p)_i \qquad [18]$$

Where coefficients $c_1$, $c_2$, and $c_3$ are computed as:

$$c_1 = \frac{1}{n} + \frac{\overline{f}(X_1,...,X_p)^2}{(n-1)\times s^2_{f(X_1,...,X_p)}}; c_2 = \frac{-\overline{f}(X_1,...,X_p)}{(n-1)\times s^2_{f(X_1,...,X_p)}}; c_3 = \frac{1}{(n-1)\times s^2_{f(X_1,...,X_p)}}$$

The leverage value has special properties that are helpful in identifying outliers. The sum of $h_{ii}$ is always equal to the number of parameters of the linear regression; in this case this value is two. Additionally, the leverage value is a measure of the distance between the $f(X_1,...,X_p)_i$ value and the mean of all model-predicted values ($\overline{f}(X_1,...,X_p)$). Therefore, the larger the leverage value the farther the i$^{th}$ data point is from the center of all $f(X_1,...,X_p)_i$ values; that is why it is called the leverage of the i$^{th}$ data point. As a rule of thumb, a data point may be considered an outlier when its leverage value is greater than $2\times p/n$ (for linear regression $p = 2$; thus greater than $4/n$).

*Studentized and semi-studentized residues*. The studentized residue is another statistic that accounts for the magnitude of the standard error of each residue (Neter et al., 1996). It is computed as shown in Equation [19]. Similarly, the semi-studentized residue can be computed based only on the MSE as shown in Equation [20].

$$\text{Studentized Residue}_i = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE \times (1 - h_{ii})}} \qquad [19]$$

$$\text{Semi-Studentized Residue}_i = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE}} \qquad [20]$$

*PRESS statistic*. The PRESS statistic is the measurement of the i$^{th}$ residue when the fitted regression is based on all of the cases except the i$^{th}$ one. This is because if $Y_i$ is far outlying, the fitted least squares regression function based on all cases including the i$^{th}$ data point value may be influenced to come close to $Y_i$, yielding a fitted value $\hat{Y}_i$ near $Y_i$. On the other hand, if the i$^{th}$ data point is excluded before the regression function is fitted, the least squares fitted value $\hat{Y}_i$ is not influenced by the outlying $Y_i$ data point, and the residual for the i$^{th}$ data point will tend to be larger and therefore more likely to expose the outlying observation. The difference between the actual observed value $Y_i$ and the estimated expected value $\hat{Y}_{i(i)}$ is computed as shown in Equation [21].

$$\text{PRESS}_i = Y_i - \hat{Y}_{i(i)} = \frac{Y_i - \hat{Y}_i}{1 - h_{ii}} \qquad [21]$$

*Studentized deleted residue*. It combines the refinements of studentized residue and PRESS calculations discussed above in diagnosing outliers or extreme data points. Equation [22] has its calculation. It is likely that the studentized deleted residue provides the most complete analysis of residues; therefore, data points with high values should be considered carefully.

$$\text{Studentized Deleted Residue}_i = \left(Y_i - \hat{Y}_i\right) \times \left[\frac{n-3}{MSE \times (n-2) \times (1-h_{ii}) - \left(Y_i - \hat{Y}_i\right)^2}\right] \qquad [22]$$

### 4.2.3. *Analysis of Residuals: Influential Points*

After identifying the possible outliers using the studentized residue techniques discussed above, the next step is to ascertain whether or not these outliers are influential.

*DFFITS influential statistic*. The DFFITS statistic is the difference between the fitted value $\hat{Y}_i$ for the i$^{th}$ data point when all n data points are used in fitting the linear regression and the predicted value $\hat{Y}_{i(i)}$ for the i$^{th}$ data point obtained when the i$^{th}$ case is omitted in fitting the linear regression. This statistic is computed as shown in Equation [23]. As a rule of thumb, a data point is potentially influential if the $|\text{DFFITS}_i|$ exceeds 1 or $2\sqrt{2/n}$ for small or large datasets, respectively.

$$\text{DFFITS}_i = \text{Studentized Deleted Residue}_i \times \sqrt{\frac{h_{ii}}{1-h_{ii}}} \qquad [23]$$

*Cook's Distance statistic*. Another measure that indicates influential patterns is the Cook's Distance (Cook's D) statistic. In contrast to DFFITS that considers the influence of the i$^{th}$ data point on the fitted value $\hat{Y}_i$, Cook's D statistic considers the influence of the i$^{th}$ data point on all *n* fitted values. Equation [24] has the calculation of Cook's D statistic.

$$\text{Cook's D}_i = \frac{Y_i - \hat{Y}_i}{2 \times MSE} \times \left(\frac{h_{ii}}{(1-h_{ii})^2}\right) \qquad [24]$$

The combined use of the residual analysis and the influential statistics provide a powerful tool for the identification of outliers or extreme data points that may affect the linear regression and therefore the evaluation of the model prediction. As a rule of thumb, when the same data point is identified by two or more residual statistics, there is a great chance that this data point is an outlier and is affecting the linear regression; however, the removal decision has to be considered with prudence.

## 4.3. *Concordance Correlation Coefficient*

Several statistics have been discussed so far that are used to assess the validity of a mathematical model but most of them give the right answer to a wrong question. The right question in assessing the model validity is if the model-predicted values are precise and accurate at the same time across a range and are tightly amalgamated along the unity line through the origin.

The concordance correlation coefficient (CCC; Lin, 1989), also known as reproducibility index, has been suggested to simultaneously account for accuracy and precision (Equation [25]). The CCC statistic is suitable for continuous variables whereas the kappa statistics is appropriate for discrete variables (Cohen, 1960, 1968).

$$\hat{\rho}_c = \frac{2 \times s_{f(X_1,\ldots,X_p)Y}}{s_Y^2 + s_{f(X_1,\ldots,X_p)}^2 + (\bar{Y} - \bar{f}(X_1,\ldots,X_p))^2} \qquad [25]$$

Equation [25] can be expanded into two components as shown in Equation [26] (Lin, 1989). The first component ($\rho$) is the correlation coefficient estimate ($r$; Equation [10]) that measures precision. The second component ($C_b$) is the bias correction factor that indicates how far the regression line deviates from the slope of unity (45°) as indicated in Equations [27] to [29].

$$\hat{\rho}_c = \hat{\rho} \times C_b \qquad [26]$$

$$C_b = \frac{2}{\left[ v + \frac{1}{v} + \mu^2 \right]} \qquad [27]$$

$$v = \frac{\sigma_1}{\sigma_2} \text{ for population}$$

$$v = \frac{s_Y}{s_{f(X_1,\ldots,X_p)}} \text{ for sample} \qquad [28]$$

$$\mu = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1 \sigma_2}} \text{ for population}$$

$$\mu = \frac{\bar{Y} - \bar{f}(X_1,\ldots,X_p)}{\sqrt{s_Y s_{f(X_1,\ldots,X_p)}}} \text{ for sample} \qquad [29]$$

The $C_b$ factor varies from 0 to 1. When $C_b$ is equal to unity, it indicates that no deviation from the unity line had occurred. However, since the $C_b$ factor does not include the $\rho$ estimate as a component, the accuracy is the same regardless of the degree of precision as long as the mean and variance are the same. The $\rho_c$ coefficient has the following properties as described by Lin (1989):

(i) $-1 \le -|\rho| \le \rho_c \le |\rho| \le 1$

(ii) $\rho_c = 0$ if and only if $\rho = 0$

(iii) $\rho_c = \rho$ if and only if $\sigma_1 = \sigma_2$ (or $s_Y = s_{f(X_1,\ldots,X_p)}$) and $\mu_1 = \mu_2$ (or $\bar{Y} = \bar{f}(X_1,\ldots,X_p)$)

(iv) $\rho_c = \pm 1$ if and only if:

a. $(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2 \times \sigma_1 \times \sigma_2 \times (1 \pm \rho) = 0$, or equivalently,

b. $\rho = \pm 1$, $\sigma_1 = \sigma_2$, and $\mu_1 = \mu_2$, or equivalently,

c. Each pair of readings is in perfect agreement or perfect reverse agreement.

The estimates $v$ and $u$ depicted in Equations [28] and [29] measure the scale shift (ratio of two standard deviations) and the location shift relative to the scale (squared difference of the means relative to the product of two standard deviations); respectively.

Assuming $\hat{\rho}_c$ is the sample CCC of paired samples from a bivariate normal distribution, its variance can be computed as described by Lin (1989) with the corrections reported by Lin (2000) (Equation [30]).

$$\sigma^2_{\hat{\rho}_c} = \frac{1}{n-2}\left[ \frac{\left(1-\rho^2\right)\rho_c^2\left(1-\rho_c^2\right)}{\rho^2} + \frac{2\rho_c^3\left(1-\rho_c\right)\mu^2}{\rho} - \frac{\rho_c^4 \mu^4}{2\rho^2} \right] \qquad [30]$$

Furthermore, Lin (1989) suggested the inverse hyperbolic tangent transformation to improve the normal approximation (Equations [31] and [32]). Monte Carlo simulation indicated that this transformation is robust for samples from the uniform and Poisson distribution even with small sample size (≤ 10 data points) (Lin, 1989). Therefore, the transformed $\rho_c$ and its variance can be computed as follows.

$$\hat{Z}_{\hat{\rho}_c} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2}\left[\ln\left(\frac{1+\hat{\rho}_c}{1-\hat{\rho}_c}\right)\right] \qquad [31]$$

$$\sigma^2_{\hat{Z}_{\hat{\rho}_c}} = \frac{\sigma^2_{\hat{\rho}_c}}{\left(1-\rho_c^2\right)^2} = \frac{1}{n\text{-}2}\left[\frac{\left(1-\rho^2\right)\rho_c^2}{\rho^2\left(1-\rho_c^2\right)} + \frac{2\rho_c^3\left(1-\rho_c\right)\mu^2}{\rho\left(1-\rho_c^2\right)^2} - \frac{\rho_c^4\mu^4}{2\rho^2\left(1-\rho_c^2\right)^2}\right] \qquad [32]$$

The confidence interval of $\rho_c$ when using the inverse hyperbolic tangent transformation is bound to the interval [-1, 1], which provides a more realistic asymmetrical interval. The un-transformed confidence interval can be computed as shown in Equation [33]. The two-tail tabular values for Z are 1.96, 1.64, and 1.28 for type I error (α/2) of 2.5%, 5%, and 10%, respectively.

$$\frac{Exp\left(2\times\left(\hat{Z}_{\hat{\rho}_c} - Z\times\sigma_{\hat{Z}}\right)-1\right)}{Exp\left(2\times\left(\hat{Z}_{\hat{\rho}_c} - Z\times\sigma_{\hat{Z}}\right)+1\right)} \le CI \le \frac{Exp\left(2\times\left(\hat{Z}_{\hat{\rho}_c} + Z\times\sigma_{\hat{Z}}\right)-1\right)}{Exp\left(2\times\left(\hat{Z}_{\hat{\rho}_c} + Z\times\sigma_{\hat{Z}}\right)+1\right)} \qquad [33]$$

This CCC statistic has been used in evaluation of measuring devices such as "gold-standard" assays, chemical method comparisons, and instrument or assay validation (Dhanoa et al., 1999; Lin, 1992; Lin et al., 2002). The departure from the true values can be separated into two components: (a.) precision is usually measured by the Person's correlation coefficient and is not correctable and (b.) instrument or method accuracy is measured by the $C_b$ coefficient and is correctable by calibration or changes in the procedures or methods. A table showing sample size needed to detect bias ($u$ and/or $v$) based on precision loss acceptability was reported by Lin (1992).

Several concerns about the unexplainable or misleading results from the $\hat{\rho}_c$ and/or $C_b$ have been discussed (Liao, 2003; Liao and Lewis, 2000; Nickerson, 1997). The CCC computed by Equation [26] assumes that each paired data point ( $f(X_1,...,X_p)_i, Y_i$ ) are interchangeable (i.e. indistinguishable without a covariance effect), although each pair of data points belongs to either one or another set of data (Nickerson, 1997). Intraclass concordance coefficients have previously been developed to estimate the degree of absolute agreement between non-interchangeable (i.e. distinguishable) measurements using either one- or two-way statistical models with fixed and/or random effects (McGraw and Wong, 1996a, b). Nickerson (1997) provided the calculation of the ICC as shown in Equation [34].

$$r_2 = \frac{2\times s_{f(X_1,...,X_p)Y}}{s_Y^2 + s^2_{f(X_1,...,X_p)} + (\bar{Y} - \bar{f}(X_1,...,X_p))^2 - \frac{s^2_{f(X_1,...,X_p)-Y}}{n}} \qquad [34]$$

The difference between Equation [25] and [34] is the term ( $s^2_{(f(X_1,...,X_p)\text{-}Y)}/n$ ) in the denominator that is the squared standard error of the differences between paired data points. When it is zero, the two coefficients (Equation [25] and [34]) are identical. The term ( $s^2_{(f(X_1,...,X_p)\text{-}Y)}/n$ ) will be zero if and only if (a.) each paired data points are identical or (b.) they differ by only a positive or negative additive constant. Additionally, as $n$ increases, the closer is the value between the coefficient $r_2$ (Equation [34]) and the coefficient $\hat{\rho}_c$ (Equation [26]).

The CCC statistic (Equation [26]) is a special case of the formula for ICC (Equation [34]), which is defined for a two-way ANOVA when the following four cases are not distinguished: (1.) random row and column effects exist but row by column interaction is assumed to be null; (2.) random row effects, fixed column interaction, but row by column interaction is assumed to be null, (3.) random row and column effects, row and column interaction exists, and (4.) random row effects, fixed column

effects, row by column interaction exists. Therefore, little advantage exists in the calculation of $\hat{\rho}_c$ over existing ICC as a way to evaluate the reproducibility index (Nickerson, 1997).

Additional drawbacks of the CCC statistic as proposed by Lin (1989) were discussed by Liao and Lewis (2000). The authors claimed the $C_b$ accuracy measure is flawed, sometimes cannot give correct accuracy information, gives unexplained results, and may lead to the selection of inaccurate and inappropriate models. For instance, let X = $Y_1$ = -2, -1, 0, 1, 2 and $Y_2$ = -2, 1, 0, -1, 2. From ($Y_1$, X), $\beta_1$ = 1, $r$ = 1, $C_b$ = 1, $r_c$ = 1, and $A_r$ = 1. From ($Y_2$, X), $\beta_1$ = 0.6, $r$ = 0.6, $C_b$ = 1, $r_c$ = 0.65, and $A_r$ = 1. Undoubtedly, $Y_1$ has a better fit than $Y_2$; however, both $C_b$ and $A_r$ statistics indicated they have the same accuracy. This is because the dataset falls into the situation where $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, as mentioned on page 12. The $A_r$ statistic has an overall improvement over $C_b$ statistic in most practical situations, but not all because both statistics share the same assumptions.

Liao (2003) indicated the CCC, as estimated by Equation [25], yields misleading coefficients because it uses the squared perpendicular distance of any paired data point without accounting for the direction of the regression line. Therefore, another data point is needed to determine the direction of the regression line. Instead of the expected squared distance, the expected square of an area was proposed. Equation [35] is suggested as a measure of the accuracy to replace Equation [27]. Further information on variance calculation is provided by Liao (2003).

$$A_\rho = \frac{4\times\left(\frac{s_{f(X_1,\ldots,X_p)}}{s_Y}\right) - \rho\times\left[1+\left(\frac{s_{f(X_1,\ldots,X_p)}}{s_Y}\right)^2\right]}{(2-\rho)\times\left[1+\left(\frac{s_{f(X_1,\ldots,X_p)}}{s_Y}\right)^2\right]\times\left(\frac{\bar{Y}-\bar{f}(X_1,\ldots,X_p)}{s_Y}\right)^2} \qquad [35]$$

In the example provided by Liao (2003), the difference in accuracy between $C_b$ and $A_\rho$ was considerably large (0.57173 and 0.20538, respectively), indicating that some models may be more or less accurate depending on the method used, which may or may not depend on the direction of the regression line.

A generalized CCC statistic has been proposed to provide a unifying approach to assess agreement among two or more measures that are either continuous or categorical in scale (King and Chinchilli, 2001). Similarly, a generalized estimating equations (GEE) approach was proposed to compute CCC via three sets of estimating equations (Barnhart and Williamson, 2001). The authors indicated the advantage of the GEE are: (a.) accommodates more than two correlated readings and tests for the equality of dependent CCC, (b.) incorporates covariates predictive of the marginal distribution, (c.) identifies covariates predictive of concordance correlation; and (d.) requires minimal assumption about distribution of the dataset.

Estimation of CCC using variance components of mixed effects model has also been investigated. This approach is valuable when comparison of more than two instruments or methods are sought or when confounding covariates are taken into account. Moreover, when the instruments or methods are considered fixed effects, the ICC and the CCC are identical (Carrasco and Jover, 2003).

Other statistical methods to assess model agreement have been discussed by Lin et al. (2002). They included extensive explanation and details of mean squared deviation (MSD), total deviation index (TDI), and coverage probability (CP) among others.

### 4.4. *Diverse Evaluation Measurements*

#### 4.4.1. *Mean Bias*

Mean bias (MB) is likely to be the oldest and most used statistic to assess model accuracy. Its calculation is based on the mean difference between observed and model-predicted values (Cochran and Cox, 1957) as indicated in Equation [36].

$$MB = \frac{\sum_{i=1}^{n} (Y_i - f(X_1, ..., X_p)_i)}{n} \qquad [36]$$

A mathematical model may have a lower MB estimate if the data points are uniformly scattered around the Y = X line; therefore, a paired *t*-test of the difference between observed and model-predicted is suggested to check whether the null hypothesis that MB equals to zero (Equation [37]).

$$t_{MB} = \frac{MB}{\sqrt{\frac{\sum_{i=1}^{n} \left( \left( Y_i - f(X_1, ..., X_p)_i \right) - MB \right)^2}{n \times (n-1)}}} \qquad [37]$$

The paired *t*-test comparing the mean value of observed and model-predicted values can fail to indicate agreement at the individual pair of data points. Thus, the paired *t*-test of the mean of the difference is preferable over the paired *t*-test of the difference of the means, since the former one is less conservative and removes any covariance between the data points. As shown by Mayer et al. (1994) and Keijnen (1987), rejection rates of the $H_0$ hypothesis are much higher when the dataset have correlated errors, even with weak correlation.

#### 4.4.2. *Mean Absolute Error*

Mean absolute error (MAE) measures the mean absolute difference between observed and model-predicted values (Byers et al., 1989; Mayer and Butler, 1993), as shown in Equation [38]. It also is commonly expressed as the mean absolute percent error as indicated in Equation [39].

$$MAE = \frac{\sum_{i=1}^{n} \left( \left| Y_i - f(X_1, ..., X_p)_i \right| \right)}{n} \qquad [38]$$

$$MA\%E = \frac{100}{n} \sum_{i=1}^{n} \left( \frac{\left( \left| Y_i - f(X_1, ..., X_p)_i \right| \right)}{\left| Y_i \right|} \right) \qquad [39]$$

The MAE is not a true deviance measure since the negative sign is removed. It simply compares the distribution of the differences against zero, which happens only if the means are identical. The lower the MAE, the more accurate is the model. A cutoff value of 10% has been suggested as an upper limit on acceptability (Kleijnen, 1987).

#### 4.4.3. *Resistant Coefficient of Determination*

Another method to estimate $r_r^2$ coefficient was proposed by Kvålseth (1985). This method uses the medians instead of means, resulting in a coefficient that is more resistant to outliers or extreme data points. Equation [40] shows the calculation of the resistant $r_r^2$.

$$r_r^2 = 1 - \left( \frac{\underset{i=1}{\overset{n}{\mathrm{M}}} \left( \left| Y_i - \hat{Y}_i \right| \right)}{\underset{i=1}{\overset{n}{\mathrm{M}}} \left( \left| Y_i - \bar{Y} \right| \right)} \right)^2 \qquad [40]$$

Where $\underset{i=1}{\overset{n}{\mathrm{M}}}$ indicates the median of the n$^{th}$ values.

The $\rho$ coefficient, an ICC, measures the linear relationship by measuring how far observations deviate from the best-fit regression. In other words, it measures the amount of overall data variation due to between-subjects variability. Therefore, it does not distinguish between situations in which observed and model-predicted values strongly agree and those in which a strong linear relationship exists but the measurements do not agree.

### *4.4.4. Modeling Efficiency*

The modeling efficiency statistic (MEF) is similar to $\rho$, which is interpreted as the proportion of variation explained by the fitted line whereas the MEF statistic is the proportion of variation explained by the line $Y = f(X_1,...,X_p)$. This statistic has been extensively used in hydrology models (Byers et al., 1989; Loague and Green, 1991; Zacharias et al., 1996), but can certainly be used in biological models. Equation [41] shows its calculation; simply, the term $\hat{Y}_i$ in Equation [10] was replaced by the term $f(X_1,...,X_p)_i$ in Equation [41].

$$MEF = \frac{\left( \sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2 - \sum_{i=1}^{n} \left( Y_i - f(X_1,...,X_p)_i \right)^2 \right)}{\sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2} = 1 - \frac{\sum_{i=1}^{n} \left( Y_i - f(X_1,...,X_p)_i \right)^2}{\sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2} \qquad [41]$$

In a perfect fit, both statistics (Equations [10] and [41]) would result in a value equal to one. The upper bound of MEF is one and the (theoretical) lower bound of MEF is negative infinity. If MEF is lower than zero, the model-predicted values are worse than the observed mean (Loague and Green, 1991). The MEF statistic may be used as a good indicator of goodness of fit (Mayer and Butler, 1993).

### *4.4.5. Coefficient of Model Determination*

The coefficient of model determination (CD) is the ratio of the total variance of observed data to the squared of the difference between model-predicted and mean of the observed data (Equation [42] ). Therefore, the CD statistic explains the proportion of the total variance of the observed values explained by the predicted data (Loague and Green, 1991). The closer to unity the better is the model predictions.

$$CD = \frac{\sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2}{\sum_{i=1}^{n} \left( f(X_1,...,X_p)_i - \bar{Y} \right)^2} \qquad [42]$$

Zacharias et al. (1996) illustrates the use of robust objective functions, which are based on the median instead of mean in Equations [41] and [42], as the best estimator of the observed data.

### *4.4.6. Other Measures*

Several other criteria for comparing mathematical models have been discussed to appraise goodness-of-fit of simulation outcome (Green and Stephenson, 1986). No single parameter was

sufficient and necessary to adequately assess the performance of the models when comparing observed and model-predicted values. Nevertheless, the authors indicated that each criterion highlighted particular aspects of the comparison. Therefore, it suggests the performance of mathematical models can be thoroughly assessed if a group of criteria is established prior the comparison of the models. More in depth discussion about model selection from a statistical point of view was presented by Neter et al. (1996) and Burnham and Anderson (2002).

### *4.5. Mean Square Error of Prediction*

The mean square error of prediction (MSEP) is probably the most common and reliable estimate to measure the predictive accuracy of a model. A comprehensive discussion of ordinary least squares technique to evaluate linear models using MSE and MSEP was presented by Bibby and Toutenburg (1977).

The MSE (Equation [11]) assesses the precision of the fitted linear regression using the difference between observed values ($Y_i$) and regression-predicted values ($\hat{Y}_i$). In contrast, MSEP (Equation [43]) consists of the difference between observed values ($Y_i$) and model-predicted values ($f(X_1,...,X_p)_i$) rather than regression-predicted value. The variance of the estimate of MSEP is given by Equation [44].

$$MSEP = \frac{\sum_{i=1}^{n}\left(Y_i - f(X_1,...,X_p)_i\right)^2}{n} \qquad [43]$$

$$\sigma^2_{MSEP} = \frac{\sum_{i=1}^{n}\left[\left(Y_i - f(X_1,...,X_p)_i\right)^2 - MSEP\right]^2}{n-1} \qquad [44]$$

Where $Y_i$ is i$^{th}$ observed value; $X_j$ are the variables used in the model to predict $Y_i$; $f(X_1,...,X_p)_i$ is the i$^{th}$ model-predicted value using $X$ variables; $\hat{Y}_i$ is the i$^{th}$ regression-predicted value of the i$^{th}$ model-predicted value using $X$ variables; MSEP is mean square error of prediction; $\sigma^2_{MSEP}$ is variance of the MSEP, and $n$ is number of data points.

Some drawbacks of this MSEP analysis are that MSEP (or its root) is a poor indicator of model performance because it removes the negative sign, weights the deviation by their squares, thus giving more influence to larger data points, and does not provide any information about model precision (Mitchell and Sheehy, 1997).

When each pair of the data ($f(X_1,...,X_p)_i, Y_i$) is mutually independent (i.e. the outcome of one pair does not depend on the outcome of another pair), and the model is independent (i.e. the parameters of the model were derived from independent experiments and were not adjusted to the current experiment been predicted), the MSEP estimate is a reliable measure of model accuracy. Nonetheless, the reliability will decrease as $n$ decreases. Therefore, the estimate of MSEP variance has an important role. Wallach and Goffinet (1987) has discussed the necessary adjustments when the pair of data is not mutually independent.

A different scenario arises when the model parameters are adjusted to the dataset. It is recognized that Equation [43] will underestimate the real MSEP because the model will reproduce more closely the data been modeled than it would for the entire population of interest. Two options are available to overcome the adjustment and evaluation conundrum (Wallach and Goffinet, 1987). Firstly, the c*ross-validation technique* is used to split the data been predicted into two groups with or without similar number of pairs of data. The first group is used for model adjustment while the second one is used for model evaluation (Picard and Cook, 1984). This technique is inappropriate for small sample

size because the variance of MSEP is likely to increase considerably when splitting the data. Secondly, the *bootstrap technique* is used to resample the same data for adjustment and for evaluation purposes. The bootstrap technique has been proven to be a good re-sampling technique (Efron, 1979, 1983) for MSEP estimation. Wallach and Goffinet (1989) have explained and well detailed a procedure to correctly estimate the population MSEP using the bootstrap method. Basically, *b* random sub-samples of size *m* are withdrawn from the original dataset, the model parameters are adjusted using the $b^{th}$ sub-sample, the difference of MSEP for the $b^{th}$ sub-sample is calculated (the MSEP should be small because the model has been adjusted for this $b^{th}$ sub-sample) and for the original sample, this difference is averaged over *b* sub-samples, and added to the MSEP of the original dataset (Wallach and Goffinet, 1987).

*Using MSEP as a direct comparison of mathematical models.* A simple approach to select for model adequacy between two (or more) models is computing the MSEP of each model and choosing the one with smaller MSEP estimates. A slightly different approach has been studied by Wallach and Goffinet (1989) that uses the ΔMSEP, difference in MSEP values between two models, as shown in Equation [45].

$$\Delta MSEP = \frac{\sum_{i=1}^{n}\left[\left(Y_i - f(X_1,...,X_p)_i\right)^2 - \left(Y_i - g(Z_1,...,Z_k)_i\right)^2\right]}{n} \qquad [45]$$

Where ΔMSEP is difference is MSEP values, $Y_i$ is $i^{th}$ observed value; *X* are the variables used in the *f(·)* model to predict the $i^{th}$ observed value; *Z* are the variables used in the *g(·)* model to predict the $i^{th}$ observed value; $f(X_1,...,X_p)_i$ is the $i^{th}$ *f(·)* model-predicted value using *X* variables; and $g(Z_1,...,Z_k)_i$ is the $i^{th}$ *g(·)* model-predicted value using *Z* variables.

If the estimated difference (ΔMSEP) is positive then model $g(Z_1,...,Z_k)_i$ is preferred, while if ΔMSEP is negative then model $f(X_1,...,X_p)_i$ is preferred since it has the smaller MSEP estimate. As usual, the variance of the ΔMSEP is of extreme importance to indicate if the ΔMSEP estimate is different from zero (reject $H_o$: ΔMSEP = 0). The calculation of the variance of ΔMSEP is shown in Equation [46].

$$\sigma^2_{\Delta MSEP} = \frac{\sum_{i=1}^{n}\left[\left\{\left(Y_i - f(X_1,...,X_p)_i\right)^2 - \left(Y_i - g(Z_1,...,Z_k)_i\right)^2\right\} - \Delta MSEP\right]}{n-1} \qquad [46]$$

If the bootstrap technique is used for data resampling, then adjustments have to be made to take into account the underprediction of MSEP when refitting the parameters of the model (Wallach and Goffinet, 1989).

*Using MSEP to evaluate model parameterization.* Bunke and Droge (1984) presented a partitioning method that takes into account the source of variation of the parameters ( $\hat{p}$ ) of the model as shown in Equations [47] to [50].

$$MSEP_{\hat{p}} = E\left\{\left[Y_i - f(X_1,...,X_p,\hat{p})_i\right]^2 \mid \hat{p}\right\} \qquad [47]$$

Decomposing $MSEP_{\hat{p}}$,

$$MSEP_{\hat{p}} = \Lambda + \Delta_{\hat{p}} \qquad [48]$$

Where,

$$\Lambda = E\left[Y_i - E\left(Y_i \mid X_j\right)\right]^2 \qquad [49]$$

$$\Delta_{\hat{p}} = E\left\{\left[E\left(Y_i \mid X_j\right) - f(X_1,...,X_p,\hat{p})_i\right]^2 \mid \hat{p}\right\} \quad [50]$$

The $\Lambda$ estimate represents the minimum $MSEP_{\hat{p}}$ that can be achieved for a set of variables $X$ and $\Delta_{\hat{p}}$ indicates the additional contribution of model bias due to the parameters of a set of variables $X$ used in the model prediction (Wallach and Goffinet, 1987). For instance, if $\Lambda$ has a greater contribution to $MSEP_{\hat{p}}$ than $\Delta_{\hat{p}}$, then other variables have to be used in the model to improve its prediction. On the other hand, if $\Delta_{\hat{p}}$ has a greater contribution to $MSEP_{\hat{p}}$ than $\Lambda$, then the model structure has to be changed. An example of application is given in Wallach and Goffinet (1987).

*Decomposing the sources of variation of MSEP*. Theil (1961) has introduced methods to decompose the sources of variation of MSEP to analyze model adequacy. Equation [43] can be expanded and solved for known measures of linear regression (Equations [51] to [52]) rather than individual pair of data.

$$MSEP = \frac{\sum_{i=1}^{n}\left[\left(\bar{f}(X_1,...,X_p) - \bar{Y}\right) + \left(f(X_1,...,X_p)_i - \bar{f}(X_1,...,X_p)\right) - \left(Y_i - \bar{Y}\right)\right]^2}{n} \quad [51]$$

Simplifying Equation [51],

$$MSEP = \left(\bar{f}(X_1,...,X_p) - \bar{Y}\right)^2 + s^2_{f(X_1,...,X_p)} + s^2_Y - 2 \times r \times s_{f(X_1,...,X_p)} \times s_Y \quad [52]$$

Theil (1961) further re-arranged Equation [52] to yield two readily interpretable forms (Equations [53] and [54]) with three distinct terms.

$$MSEP_1 = \left(\bar{f}(X_1,...,X_p) - \bar{Y}\right)^2 + \left(s_{f(X_1,...,X_p)} - s_Y\right)^2 + 2 \times (1 - r) \times s_{f(X_1,...,X_p)} \times s_Y \quad [53]$$

$$MSEP_2 = \left(\bar{f}(X_1,...,X_p) - \bar{Y}\right)^2 + \left(s_{f(X_1,...,X_p)} - r \times s_Y\right)^2 + \left(1 - r^2\right) \times s^2_Y \quad [54]$$

Where MSEP is mean square error of prediction; $f(X_1,...,X_p)_i$ is the i$^{th}$ model-predicted value using variables $X$; $Y_i$ is the i$^{th}$ observed value; $s^2$ (and $s$) are the variance (and standard deviation) associated with observed and model-predicted values; and $r$ is the coefficient of correlation. Note that $s^2$ (and $s$) are associated with $n$ observations and not ($n$ – 1) degrees of freedom as computed by Equations [7] and [8].

The three terms in Equation [53] may be interpreted as errors in central tendency (or mean shift), errors due to unequal variation (variances), and errors due to incomplete (co)variation, respectively. In short, these terms measure the mean bias, variance, and (co)variance. Similarly, the three terms in Equation [54] represent errors in central tendency, errors due to regression, and errors due to disturbances (or random errors), i.e. unexplained variance that cannot be accounted for by the linear regression. Attention should be drawn to the later error term since it might be a good indicator of lack-of-fit.

Furthermore, the second term of Equation [54] can be re-written to asses the variability of the slope from unity as shown in Equations [55] to [59].

$$\left(s_{f(X_1,...,X_p)} - r \times s_Y\right)^2 = s^2_{f(X_1,...,X_p)} - 2 \times s_{f(X_1,...,X_p)} \times r \times s_Y + r^2 \times s^2_Y \quad [55]$$

Substituting $r$ by Equation [10] and simplifying,

$$\left(s_{f(X_1,\ldots,X_p)} - r \times s_Y\right)^2 = s^2_{f(X_1,\ldots,X_p)} - \frac{2 \times \sum_{i=1}^{n}\left[\left(Y_i - \overline{Y}\right) \times \left(f(X_1,\ldots,X_p)_i - \overline{f}(X_1,\ldots,X_p)\right)\right]}{n} +$$

$$+ \frac{\left\{\sum_{i=1}^{n}\left[\left(Y_i - \overline{Y}\right) \times \left(f(X_1,\ldots,X_p)_i - \overline{f}(X_1,\ldots,X_p)\right)\right]\right\}^2}{n^2 \times s^2_{f(X_1,\ldots,X_p)}} \qquad [56]$$

Using Equation [5] and solving for b,

$$\left(s_{f(X_1,\ldots,X_p)} - r \times s_Y\right)^2 = s^2_{f(X_1,\ldots,X_p)} - \frac{2 \times b \times \sum_{i=1}^{n}\left(f(X_1,\ldots,X_p)_i - \overline{f}(X_1,\ldots,X_p)\right)^2}{n} +$$

$$+ \frac{b^2 \times \left[\sum_{i=1}^{n}\left(f(X_1,\ldots,X_p)_i - \overline{f}(X_1,\ldots,X_p)\right)^2\right]^2}{n^2 \times s^2_{f(X_1,\ldots,X_p)}} \qquad [57]$$

Using Equation [8] and solving for $s^2_{f(X_1,\ldots,X_p)}$,

$$\left(s_{f(X_1,\ldots,X_p)} - r \times s_Y\right)^2 = s^2_{f(X_1,\ldots,X_p)} - \frac{2 \times b \times s^2_{f(X_1,\ldots,X_p)} \times n}{n} + \frac{b^2 \times \left(s^2_{f(X_1,\ldots,X_p)}\right)^2 \times n^2}{n^2 \times s^2_{f(X_1,\ldots,X_p)}} \qquad [58]$$

Simplifying and solving for *b*,

$$\left(s_{f(X_1,\ldots,X_p)} - r \times s_Y\right)^2 = s^2_{f(X_1,\ldots,X_p)} \times \left(1 - 2 \times b + b^2\right) = s^2_{f(X_1,\ldots,X_p)} \times (1-b)^2 \qquad [59]$$

Therefore, Equation [54] can be re-written as in an equivalent form to represent the linear regression slope as shown in Equation [60].

$$MSEP_3 = \left(\overline{f}(X_1,\ldots,X_p) - \overline{Y}\right)^2 + s^2_{f(X_1,\ldots,X_p)} \times (1-b)^2 + \left(1 - r^2\right) \times s^2_Y \qquad [60]$$

The terms from Equations [53], [54] (or [60]) are usually divided by the total MSEP to obtain the inequality proportions of MSEP (Theil, 1961). This approach facilitates the identification of areas that need improvement or are affected by certain changes in the model. Table 3 identifies the five relevant inequality proportions.

Table 3. Description of equations used to compute inequality proportions [a]

| Inequality Proportions | Equations | Descriptions |
|---|---|---|
| $U^M$ | $\left(\overline{f}(X_1,\ldots,X_p) - \overline{Y}\right)^2 / MSEP$ | Mean bias |
| $U^S$ | $\left(s_{f(X_1,\ldots,X_p)} - s_Y\right)^2 / MSEP$ | Unequal variances |
| $U^C$ | $2 \times (1-r) \times s_{f(X_1,\ldots,X_p)} \times s_Y / MSEP$ | Incomplete (co)variation |
| $U^R$ | $s^2_{f(X_1,\ldots,X_p)} \times (1-b)^2 / MSEP$ | Systematic or slope bias |
| $U^D$ | $\left(1 - r^2\right) \times s^2_Y / MSEP$ | Random errors |

[a] Note that $U^M + U^S + U^C = U^M + U^R + U^D = 1$

Applications of the MSEP decomposition have been reported in the literature for comparing accuracy among models (Gauch et al., 2003; Kobayashi and Salam, 2000; Kohn et al., 1998; Traxler et al., 1998), model adequacy (Tedeschi et al., 2000), and instrument calibration (Dhanoa et al., 1999).

Table 4 contains the comparison of inequality proportions of the case scenarios depicted in Figure 2. Case 1 shows a combination of inaccuracy (bias) and imprecision (uncertainty) in which more than 80% of the error is due to random error or unexplained variation by the model. As expected, mean bias is the only source of variation in case 2 whereas random error is the only source of variation in case 3.

Table 4. Comparison of inequality proportions of the case scenarios from Figure 2 [a]

| Inequality proportions | Case scenarios | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| MSEP | 1.67 | 1 | 0.5 | 0 |
| Mean bias ($U^M$), % | 6.67 | 100 | 0 | 0 |
| Unequal variances ($U^S$), % | 11.1 | 0 | 13.9 | 0 |
| Incomplete (co)variation ($U^C$), % | 82.2 | 0 | 86.1 | 0 |
| Systematic or slope bias ($U^R$), % | 10.0 | 0 | 0 | 0 |
| Random errors ($U^D$), % | 83.3 | 0 | 100 | 0 |

[a] Note that $U^M + U^S + U^C = U^M + U^R + U^D = 100\%$

### *4.6. Nonparametric Analysis*

The assessment of adequacy of a mathematical model may often be related to its ability to yield the same ranking between observed and model-predicted values rather than model-predicted on observed values per se. For instance, one might be interested in testing if the outcomes of a mathematical model ranks the most efficient bull or highest crop yield accordingly to the rank of the real system, regardless its value of efficiency or productivity per se. Nonparametric tests are more resilient to abnormalities of the data such as outliers; therefore, it may generally be used to check model adequacy.

#### *4.6.1. Spearman Correlation*

Similar to Equation [10], the Spearman correlation ($r_S$), also known as the linear correlation coefficient of the ranks, replaces each data point by its rank among all other data points. Nonparametric correlation is more robust than parametric correlation ($\rho$), being more resistant to unplanned defects in the data outlier. Equation [61] shows the calculation of $r_S$ coefficient. The Spearman correlation is closely inversely related to the sum of squared differences of ranks (Equation [62]) (Agresti, 1996, 2002; Press et al., 1992).

$$r_S = \frac{\sum_{i=1}^{n}\left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_i - \bar{R}\right)^2}\sqrt{\sum_{i=1}^{n}\left(S_i - \bar{S}\right)^2}} \qquad [61]$$

$$D = \sum_{i=1}^{n}\left(R_i - S_i\right)^2 \qquad [62]$$

Where $R_i$ and $S_i$ are the rank of the i$^{th}$ model-predicted and observed value.

### 4.6.2. *Kendall's Coefficient*

Similar to the CCC statistic, Kendall's τ measures the ordinal concordance of $\frac{1}{2}\times n\times(n-1)$ data points where a data point cannot be paired with itself and either order count as one pair. A data pair is concordance when its relative ordering of the ranks of the two X's is the same as the relative ordering of the rank of their respective Y's. A data pair is discordance when its relative ordering of the ranks of the two X's is opposite from the relative ordering of the rank of their respective Y's. When a tie occurs either between X's or Y's, the pair is neither concordant nor discordant, but a point is assigned to Extra X or Extra Y variables; respectively. When a tie occurs on both at the same time, nothing is done. Equation [63] shows the computation of Kendall's τ statistic.

$$\tau = \frac{Concordant - Discordant}{\sqrt{Concordant + Discordant - ExtraY} \times \sqrt{Concordant + Discordant - ExtraX}} \qquad [63]$$

### 4.6.3. *Balance Analysis*

A novel technique using nonparametric tests is proposed to evaluate the balance of number of data points under- and overpredicted by the mathematical model above and below the observed or model-predicted mean. As shown in Table 5, the number of data points that occurred in each plot quadrant is represented by $n_{ij}$.

Table 5. Contingency table for data points under- and overpredicted above and below the observed/model-predicted mean

| Model prediction | Observed or Model-Predicted Mean | |
|---|---|---|
| | Below | Above |
| Overpredicted | $n_{11}$ | $n_{12}$ |
| Underpredicted | $n_{21}$ | $n_{22}$ |

Two $\chi^2$ tests are performed with the number listed in the contingency table (Table 5) as described by Agresti (1996; 2002). The first $\chi^2$ test evaluates whether the data points were homogenously distributed; that means, if 25% of the data points were allocated in each plot quadrant as shown in Equation [64]. This test indicates if the model prediction is impartial regarding the observed and model-predicted mean; that means if there is any tendency of over- or underprediction below or above the mean. The second $\chi^2$ test checks to determine whether there is any association among model prediction behavior and locations relative to the mean as shown in Equation [65]. This test indicates if the same trend of under- or overprediction is similar between the locations relative to the observed or model-predicted means. That means, if the model tends to over- or under-predict as observed or model-predicted values increase/decrease.

$$\chi_1^2 = \sum_{i=1,j=1}^{2,2} \frac{\left(n_{ij} - 0.25\times n\right)^2}{0.25\times n} \qquad [64]$$

$$\chi_2^2 = \sum_{i=1,j=1}^{2,2} \frac{\left(n_{ij} - w_{ij}\right)^2}{w_{ij}} \qquad [65]$$

$$w_{ij} = \frac{\left(n_{i.}\times n_{.j}\right)}{n} \qquad [66]$$

Where $n$ is the number of data points, $w_{ij}$ is the expected number of data points, $n_{i.}$ is the number of points in row $i$ and $n_{.j}$ is the number of points in column $j$.

Additionally, the Cramer's V and contingency coefficient C is computed as suggested by Press et al. (1992) and shown in Equations [67] and [68], respectively. Both statistics vary between zero and unity, indicating no association or perfect association, respectively.

$$V = \sqrt{\frac{\chi^2}{n \times Min(i-1, j-1)}} = \sqrt{\frac{\chi^2}{n}} \qquad [67]$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \qquad [68]$$

Where $n$ is the number of data points.

Finally, the odds ratio statistics (Agresti, 1996, 2002) is used to measure the association of the outcome of Table 5 as shown in Equation [69]. The odds ratio can equal any nonnegative number. When the variables (above or below the mean and under-or overprediction) are independent (not associated) $\hat{\theta}$ is equal to 1. When $1 < \hat{\theta} < 0$, the odds of having more overprediction is higher than underprediction below than above the mean. Because of the skewness, statistical inference if performed with the $Ln(\hat{\theta})$. The asymptotic standard error (ASE) is computed for the $Ln(\hat{\theta})$ as shown in Equation [70].

$$\hat{\theta} = \frac{(n_{11} + 0.5) \times (n_{22} + 0.5)}{(n_{12} + 0.5) \times (n_{21} + 0.5)} \qquad [69]$$

$$ASE(Ln(\hat{\theta})) = \sqrt{\frac{1}{(n_{11} + 0.5)} + \frac{1}{(n_{12} + 0.5)} + \frac{1}{(n_{21} + 0.5)} + \frac{1}{(n_{22} + 0.5)}} \qquad [70]$$

The normal distribution is used to assess the probability of $Ln(\hat{\theta})$ being different than zero, which is equivalent to having $\hat{\theta}$ different than unity. The closer $\hat{\theta}$ is to unity the lesser is the association between over- or underprediction and location of the observed or model-predicted mean, indicating the mathematical model is more or less sensitive to increases or decreases in the observed or model-predicted data points.

### *4.7. Data Distribution Comparison*

The comparison of the distribution of the observed and model-predicted values has also been utilized to identify model adequacy for stochastic (Reynolds and Deaton, 1982) and deterministic models (Dent and Blackie, 1979). The Kolmogoroff-Smirnov's D test (Kolmogoroff, 1933; Smirnov, 1933) has been used to assess the probability that two data sets (observed and model-predicted values) have the same distribution. It consists to measure the overall difference of the area between two cumulative distribution functions (Press et al., 1992) as shown in Equation [71].

$$D = \underset{-\infty < J < \infty}{Max} \left| S_{N1}(J) - S_{N2}(J) \right| \qquad [71]$$

Where $S_{N1}(J)$ and $S_{N2}(J)$ are two different cumulative distribution functions.

Analysis of graphical representation may help in deciding whether two distributions are similar between themselves. Figure 3 compares the distribution of observed dry matter intake and model-predicted dry matter required of the growth model developed by Tedeschi et al. (2004). As determined by @Risk 4.5 (Palisade Co, Newfield, NY), both follow the log-logistic distribution based on $\chi^2$ criterion.

In a stricter sense, a comparison of two distributions can be done with a normal distribution test as that suggested by Shapiro and Wilk (1965). If both datasets are deemed to be normally distributed, then a comparison of between their mean and variance will identity if the observed and model-predicted data are similar. Equation [72] shows the calculation of Shapiro-Wilk's W test.

$$W = \frac{\left(\sum_{i=1}^{n} a_i J_i\right)^2}{\sum_{i=1}^{n}\left(J_1 - \bar{J}\right)^2} \qquad [72]$$

Where $a_i$ are constants generated from means, variances, and (co)variances of the order statistics of a sample of size $n$ from a normal distribution and $J_i$ are the ordered sample values.

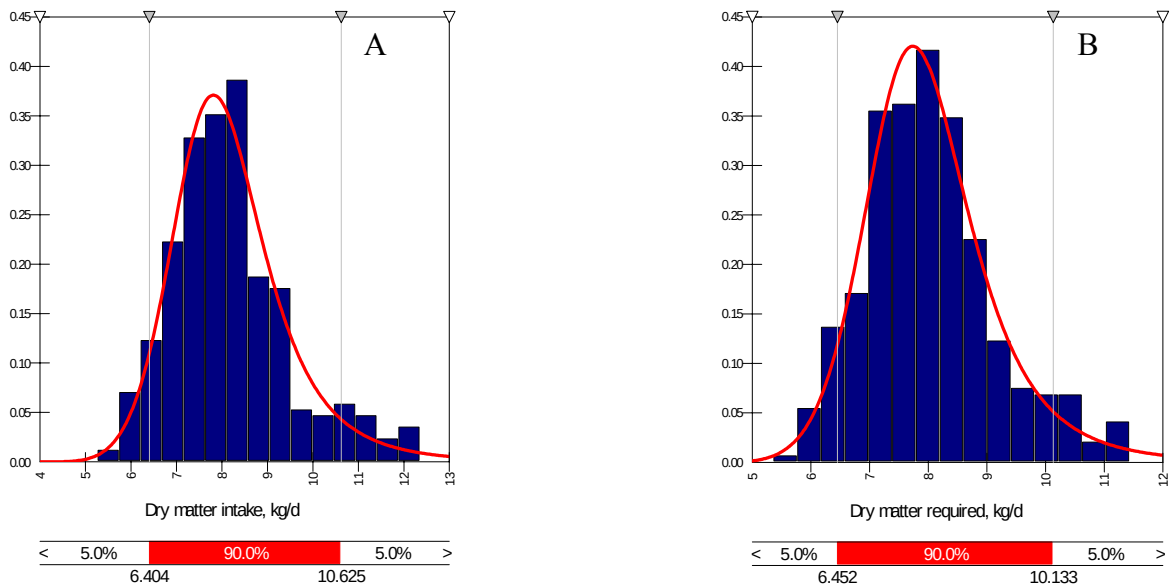



Figure 3. Distribution comparison between observed dry matter intake (A) and model-predicted dry matter required (B) from Tedeschi et al. (2004). Both observed and model-predicted values follow a log-logistic distribution determined by @Risk 4.5 (Palisade Co, Newfield, NY).

## 5. Example of Application

Two hypothetical model-predicted values were used to apply the techniques discussed above in selecting for model adequacy (Table 6). Table 7 lists some of the result of the analysis. Model B was more precise than model A ($r^2$ of 0.949 and 0.683; respectively) in accounting for the variation of the observed values. This result was confirmed by the resistant $r^2$ and MSE statistics (Table 7).

Even though the intercept was not different from zero and slope was not different from unity for both models ($P > 0.30$), the simultaneous rejected the null hypothesis for model B ($P < 0.01$) likely due to the scatter around the line for model A. The concordance correlation coefficient analysis indicated that both models had high accuracy (0.984 versus 0.945), but overall concordance correlation was smaller for model A than model B (0.813 versus 0.920; respectively) due to the smaller correlation coefficient of model A compared to model B. The $\hat{\gamma}_\rho$ estimate (Liao, 2003) had the same pattern too (not shown), suggesting that model B had a better concordance with the observed data.

Table 6. Hypothetical dataset for model adequacy comparison

| N | Observed (Y) | Model-Predicted (X) | |
|---|---|---|---|
| | | A | B |
| 1 | 1.5 | 2.05 | 1.78 |
| 2 | 2.7 | 6.48 | 4.59 |
| 3 | 3.2 | 2.50 | 4.93 |
| 4 | 4.8 | 4.97 | 4.88 |
| 5 | 5.7 | 6.00 | 6.85 |
| 6 | 6.4 | 9.09 | 7.75 |
| 7 | 8.5 | 6.50 | 10.19 |
| 8 | 8.8 | 9.45 | 9.13 |
| 9 | 9.1 | 9.91 | 9.50 |
| 10 | 10.5 | 9.00 | 12.20 |

Nonetheless, the mean bias analysis indicated that model A was more accurate with lower absolute mean bias (0.475 versus 1.06) and as a percentage of model-predicted values (7.2 versus 14.76), though both models underpredicted the observed data. Similar to the coefficient of determination, the MEF was greater for model B than model A (0.813 versus 0.648; respectively). The CD statistic was similar between models.

Table 7. Results of the comparison of observed and two model-predicted values

| Item | Equations | Model | |
|---|---|---|---|
| | | A | B |
| Linear regression | [1] | Y = 0.199 + 0.898×X | Y = -0.654 + 0.943×X |
| $r^2$ | [10] | 0.683 | 0.949 |
| Resistant $r^2$ | [40] | 0.788 | 0.928 |
| MSE | [11] | 3.345 | 0.536 |
| $P_{(\beta_0=0,\beta_1=1)}$ | [15] | 0.653 | 0.005 |
| $P_{(\beta_0=0,\beta_1=1)}$ | [16] | 0.709 | 0.010 |
| Concordance coefficient | | | |
| $C_b$ | [27] | 0.984 | 0.945 |
| $\hat{\rho}_c$ | [26] | 0.813 | 0.920 |
| $r_2$ | [34] | 0.828 | 0.923 |
| $A_\rho$ | [35] | 0.983 | 0.946 |
| Mean bias | [36] | -0.475 | -1.06 |
| Mean bias, % of X | | -7.202 | -14.76 |
| $P_{(MB = 0)}$ | [37] | 0.413 | 0.001 |
| MEF | [41] | 0.648 | 0.813 |
| CD | [42] | 1.144 | 0.834 |
| MSEP | [43] | 2.976 | 1.581 |
| MSEP decomposition | | | |
| Mean bias, $U^M$ | Table 3 | 7.58 | 71.7 |
| Slope bias, $U^R$ | Table 3 | 2.51 | 1.83 |
| Random errors, $U^D$ | Table 3 | 89.9 | 27.1 |

The MSEP for model B was lower than model A indicating a better prediction. The decomposition of the MSEP indicated different patterns in the errors of prediction between model A

and B. For model A, the lack of correlation of random errors accounted for nearly 90% of the errors whereas for model B, mean bias accounted for more than 70% of the errors of prediction.

In summary, these results indicated that model B was more precise but less accurate than model A. Selection for model adequacy requires a thorough comparison of several techniques to adequately select a model. The use of single techniques may lead to erroneous selection (e.g. mean bias).

## 6. Conclusions

The identification and acceptance of wrongness of a model is an important step towards the development of more reliable and accurate models. The assessment of the adequacy of mathematical models is only possible through the combination of several statistical analyses and proper investigation regarding the purposes for which the mathematical model was initially conceptualized and developed for. The usefulness of a model should be assessed through its sustainability for a particular purpose.

## 7. Literature Cited


Agresti, A. 1996. An introduction to categorical data analysis. Wiley-Interscience, New York.

Agresti, A. 2002. Categorical data analysis (2nd. ed.). John Wiley & Sons, New York.

Analla, M. 1998. Model validation through the linear regression fit to actual versus predicted values. Agric. Syst. 57:115-119.

Barnhart, H. X. and J. M. Williamson. 2001. Modeling concordance correlation via GEE to evaluate reproducibility. Biometrics. 57:931-940.

Bibby, J. and H. Toutenburg. 1977. Prediction and improved estimation in linear models. John Wiley & Sons, Berlin, Germany.

Box, G. E. P. 1979. Rubustness in Statistics. Academy Press, London.

Bunke, O. and B. Droge. 1984. Estimators of the mean squared error of prediction in linear regression. Technometrics. 26:145-155.

Burnham, K. P. and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.). Springer-Verlag, New York.

Byers, F. M., G. T. Schelling, and L. W. Greene. 1989. Development of growth functions to describe energy density of growth in beef cattle. Pages 195-198 in Proceedings of Energy Metabolism of Farm Animals, 11, Lunteren. Pudoc.

Carrasco, J. L. and L. Jover. 2003. Estimating the generalized concordance correlation coefficient throught variance components. Biometrics. 59:849-858.

Cochran, W. G. and G. M. Cox. 1957. Experimental Design. John Wiley & Sons, New York.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20:37-46.

Cohen, J. 1968. Weighted kappa: Normal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 70:213-220.

Dent, J. B. and M. J. Blackie. 1979. Systems Simulation in Agriculture. Applied Science, London.

Dhanoa, M. S., S. J. Lister, J. France, and R. L. Barnes. 1999. Use of mean square prediction error analysis and reproducibility measures to study near infrared calibration equation performance. J. Near Infrared Sprectrosc. 7:133-143.

Efron, B. 1979. Bootstrap methods: Another data look at the jacknife. The annals of Statistics. 7:1-26.

Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of American Statistical Association. 78:316-331.

Forrester, J. W. 1961. Industrial Dynamics. MIT Press, Cambridge.

Gauch, H. G., J. T. G. Hwang, and G. W. Fick. 2003. Model evaluation by comparison of model-based predictions and measured values. Agron. J. 95:1442-1446.

Green, I. R. A. and D. Stephenson. 1986. Criteria for comparison of single event models. Journal of Hydrology Sciences. 31:395-411.

Gunst, R. F. and R. L. Mason. 1980. Regression analysis and its application. Marcel Dekker, New York, NY.
Hamilton, M. A. 1991. Model validation: an annotated bibliography. Communications in Statistics: Theory & Methods. 20:2207-2266.
Harrison, S. R. 1990. Regression of a model on real-system output: an invalid test of model validity. Agric. Syst. 34:183-190.
Harrison, S. R. 1991. Validation of agricultural expert systems. Agric. Syst. 35:265-285.
Jefferys, W. H. 1980. On the method of least squares. Astron. J. 85:177-181.
Jefferys, W. H. 1981. On the method of least squares. II. Astron. J. 86:149-155.
Jefferys, W. H. 1988a. Erratum: "On the method of least squares". Astron. J. 95:1299.
Jefferys, W. H. 1988b. Erratum: "On the method of least squares. II". Astron. J. 95:1300.
King, T. S. and V. M. Chinchilli. 2001. A generalized concordance correlation coefficient for continuous and categorical data. Statist. Med. 20:2131-2147.
Kleijnen, J. P. C. 1987. Statistical Tools for Simulation Practitioners. Marcel Dekker, New York, NY.
Kleijnen, J. P. C., B. Bettonvil, and W. Van Groenendall. 1998. Validation of trace-driven simulation models: a novel regression test. Management Science. 44:812-819.
Kobayashi, K. and M. U. Salam. 2000. Comparing simulated and measured values using mean squared deviation and its components. Agron. J. 92:345-352.
Kohn, R. A., K. F. Kalscheur, and M. Hanigan. 1998. Evaluation of models for balancing the protein requirements of dairy cows. J. Dairy Sci. 81:3402-3414.
Kolmogoroff, A. N. 1933. Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari. 4:83-91.
Kvålseth, T. O. 1985. Cautionary note about $R^2$. The American Statistician. 39:279-285.
Liao, J. J. Z. 2003. An improved concordance correlation coefficient. Pharmaceut. Statist. 2:253-261.
Liao, J. J. Z. and J. Lewis. 2000. A note on concordance correlation coefficient. PDA Journal of Pharmaceutical Science and Technology. 54:23-26.
Lin, L. I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 45:255-268.
Lin, L. I.-K. 1992. Assay validation using the concordance correlation coeficient. Biometrics. 48:599-604.
Lin, L. I.-K. 2000. A note on the concordance correlation coefficient. Biometrics. 56:324-325.
Lin, L. I.-K., A. S. Hedayat, B. Sinha, and M. Yang. 2002. Statistical methods in assessing agreement: Models, issues, and tools. J. Am. Statist. Assoc. 97:257-270.
Loague, K. and R. E. Green. 1991. Statistical and graphical methods for evaluating solute transport models: Overviewa and application. Journal of Contaminant Hydrology. 7:51-73.
Lybanon, M. 1984. A better least-squares method when both variables have uncertainties. Am. J. Phys. 52:22-26.
Mayer, D. G. and D. G. Butler. 1993. Statistical validation. Ecol. Model. 68:21-32.
Mayer, D. G., M. A. Stuart, and A. J. Swain. 1994. Regression of real-world data on model output: an appropriate overall test of validity. Agric. Syst. 45:93-104.
McGraw, K. O. and S. P. Wong. 1996a. Correction to "Forming inferences about some intraclass correlation coefficients". Psychological Methods. 1:390.
McGraw, K. O. and S. P. Wong. 1996b. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1:30-46.
Mitchell, P. L. 1997. Misuse of regression for empirical validation of models. Agric. Syst. 54:313-326.
Mitchell, P. L. and J. E. Sheehy. 1997. Comparison of predictions and observations to assess model performance: a method of empirical validation. Pages 437-451 in Applications of systems approaches at the field level. M. J. Kropff, P. S. Teng, P. K. Aggarwal, J. Bouma, B. A. M. Bouman, J. W. Jones and H. H. Van Laar, ed. Kluwer Academic, Boston, MA.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied Linear Statistical Models (4th ed.). McGraw-Hill Publishing Co., Boston.

Nickerson, C. A. E. 1997. A note on "A concordance correlation coefficient to evaluate reproducibility". Biometrics. 53:1503-1507.

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science. 263:641-646.

Picard, R. R. and R. D. Cook. 1984. Cross-validation of regression models. J. Am. Stat. Assoc. 79:575-583.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. Numerical Recipes in Fortran 77: The art of scientific computing (2nd ed.). Cambridge University Press.

Reed, B. C. 1989. Linear least-squares fits with errors in both coordinates. Am. J. Phys. 57:642-646.

Reed, B. C. 1990. Erratum: "Linear least-squares fits with errors in both coordinates". Am. J. Phys. 58:189.

Reed, B. C. 1992. Linear least-squares fits with errors in both coordinates. II: Comments on parameter variances. Am. J. Phys. 60:59-62.

Reynolds, J., M.R. and M. L. Deaton. 1982. Comparison of some tests for validation of stochastic simulation models. Comm. Stat. Simul. Comput. 11:769-799.

Shaeffer, D. L. 1980. A model evaluation methodology applicable to environmental assessment models. Ecol. Model. 8:275-295.

Shapiro, S. S. and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). Biometrika. 52:591-611.

Smirnov, N. 1933. Estimate of deviation between empirical distribution functions in two independent samples. Moscow University Mathematics Bulletin. 2:3-16.

Squire, P. T. 1990. Comment on "Linear least-squares fits with errors in both coordinates". Am. J. Phys. 58:1209.

Sterman, J. D. 2000. Business Dynamics: Systems thinking and modeling for a complex world. Irwin McGraw-Hill, New York.

Sterman, J. D. 2002. All models are wrong: reflections on becoming a system scientist. System Dynamics Review. 18:501-531.

Tedeschi, L. O., D. G. Fox, and P. J. Guiroy. 2004. A decision support system to improve individual cattle management. 1. A mechanistic, dynamic model for animal growth. Agric. Syst. 79:171-204.

Tedeschi, L. O., D. G. Fox, and J. B. Russell. 2000. Accounting for the effects of a ruminal nitrogen deficiency within the structure of the Cornell net carbohydrate and protein system. J. Anim. Sci. 78:1648-1658.

Theil, H. 1961. Economic forecasts and policy. Pages 6-48 in Contributions to Economic Analysis. R. Strotz, J. Tinbergen, P. J. Verdoorn and H. J. Witteveen, ed. (2nd ed.) North-Holland Publishing Company, Amsterdam.

Traxler, M. J., D. G. Fox, P. J. Van Soest, A. N. Pell, C. E. Lascano, D. P. D. Lanna, J. E. Moore, R. P. Lana, M. Vélez, and A. Flores. 1998. Predicting forage indigestible NDF from lignin concentration. J. Anim. Sci. 76:1469-1480.

Wallach, D. and B. Goffinet. 1987. Mean square error of prediction in models for studying ecological and agronomic systems. Biometrics. 43:561-573.

Wallach, D. and B. Goffinet. 1989. Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecol. Model. 44:299-306.

Waller, L. A., D. Smith, J. E. Childs, and L. A. Real. 2003. Monte Carlo assessments of goodness-of-fit for ecological simulation models. Ecol. Model. 164:49-63.

Zacharias, S., C. D. Heatwole, and C. W. Coakley. 1996. Robust quantitative techniques for validating pesticide transport models. Trans. ASAE. 39:47-54.