ON USING MODEL PERFORMANCE STATISTICS IN APPLYING MODELS

Akula Venkatram and Wenjun Qian

Mechanical Engineering, University of California, Riverside, CA

Abstract: This paper makes the case for developing a statistical model to describe the behavior of the residuals between model estimates and corresponding observations. Using a framework that relates model estimates to corresponding observations, we show that the distribution of the residuals can be conveniently characterized by the geometric mean, m_g, and the geometric standard deviation, s_g, of the ratio of the observed to the model estimates of the variable of interest. We demonstrate the role of these statistics in the application of the model. Postulating a linear relationship between the residual and the model estimate allows us to separate the residual/model error into two components: one that is correlated to the model estimate and can be thus reduced in principle through model improvement, and a component that can be reduced only by expanding the model input set. The second part of the paper incorporates this description of model error into a graphical representation that builds upon that proposed by Taylor, A. (2001).

***Key words**: Model performance, performance statistics, Taylor diagram, model evaluation, inherent error, model application.*

1. INTRODUCTION

A wide variety of statistics is used to judge the performance of regional air quality models (Chang, J.C. and S. R. Hanna, 2004). The most commonly cited statistics are the normalized bias, *D*, and the normalized gross error, E_d, which are defined by

$$D = \left\langle \left(\frac{C_p - C_o}{C_o} \right) \right\rangle$$
$$E_d = \left\langle \left(\frac{\left| C_p - C_o \right|}{C_o} \right) \right\rangle, \qquad (1)$$

where C_p is the model estimate, C_o is the corresponding observed value, and the angle brackets refer to averages over time at a single location or over spatial locations at a single time. The estimated and observed concentrations can also correspond to maximum values in a 24 hour period, and might not refer to the same location or time.

Model performance statistics are generally used to compare the relative performances of several models or to accept or reject a model. For example, the current consensus is that a photochemical model is acceptable if maximum daily ozone estimates yield a bias $D \le \pm 0.15$, and a gross error $E_d \le 0.35$. These limits are based on experience with existing air quality models.

In this note, we demonstrate the need to model these statistics explicitly, and use them in applying the model. Then the model estimate consists of two components: a deterministic estimate based on known inputs, and a statistical estimate based on the model for the residuals between model estimates and observations. Both components play a role in describing observations. The next section presents the formal framework that allows formulation of the model for the residuals between model estimates and observations. In the second part of this paper, we show how the results from this model can be incorporated into a graphical representation, which builds upon that proposed by Taylor, A. (2001). The ideas proposed in this paper are illustrated through examples.

2. MODEL EVALUATION FRAMEWORK

A model prediction will always differ from the corresponding observation because the model cannot include all the variables that affect the observation. The best that the model can do is to provide an estimate of the average over the ensemble of all possible observations corresponding to the model inputs, α (Venkatram, A. 1982; Weil J. et al., 1992). Because observations respond to a set β not included in the model, we have an infinite ensemble of observations associated with a given model input set α. Then, we can write

$$C_o(\alpha,\beta) = C_p(\alpha) + \varepsilon(\alpha,\beta) \qquad (2)$$

where *C* refers to the variable of interest, such as concentration, the subscript *o* refers to an observation, and *p* refers to the model estimate or prediction. The residual between model prediction and observation, $\varepsilon(\alpha,\beta)$, is the error associated with the lack of knowledge of β, and is inherent in the sense that it can be decreased only by increasing the set α to include more of β. In addition, there are other errors related to model inputs and model formulation that can reduced without changing α. The analysis proposed in this paper attempts to separate these two types of error.

The statistics of residuals between model estimates and observations can be used to construct “skill” scores to rank the ability of models to explain observations. However, these statistics serve another important function, which is to serve as a necessary nondeterministic component of the model estimate that needs to be explicitly considered in the application of the model. The next section discusses this role of residual statistics.

3. COMPUTING AND APPLYING MODEL PERFORMANCE STATISTICS

We rewrite the relationship between an estimate, C_p, and the associated infinite set of observed values, C_o:

$$\hat{C}_o = \hat{C}_p + \varepsilon \ , \tag{3}$$

where the symbol '^' denotes some transformation of the observed and the predicted value that allows convenient analysis of the error. The error, ε, now includes errors in model inputs, errors in model formulation, and variations of factors not included in the model. Calculating the statistics of is not a straightforward manipulation of the residuals between model estimates and corresponding observations. This is because the residual between model prediction and observation depends on the definition of the ensemble in terms of the model input set α. In practice, we cannot keep α fixed and conduct a set of experiments to obtain the residuals required to calculate the statistics. All we usually have is a single residual for an ensemble corresponding to a model input set. This means that before we can calculate residual statistics, we have to either check whether the residuals are independent of model inputs or transform them to make them independent. The analysis assumes that an appropriate transformation has been used to make the standard deviation of the residuals independent of the model estimate.

There is some empirical evidence and theoretical justification (See Csanady, 1973) for describing observed concentrations with the lognormal distribution. Then it is reasonable to assume that observed concentrations are log-normally distributed about the model prediction so that residual statistics can be calculated from the equation:

$$\ln(C_o) = \ln(C_p) + \varepsilon \quad , \tag{4}$$

The distribution of these residuals can be is conveniently characterized in terms of the geometric mean and the standard deviation:

$$\begin{aligned} &m_g = \exp\left(mean \text{ of } \varepsilon\right) \\ &and \qquad\qquad\qquad\qquad , \\ &s_g = \exp\left(s\tan dard \text{ deviation of } \varepsilon\right) \end{aligned} \tag{5}$$

where the angle brackets refer to an average. We see that that the deviation of the geometric mean, m$_g$, from unity tells us whether the model is underpredicting or overpredicting; it is a measure of bias of the model estimate.

The calculation of the geometric mean, m_g, and the geometric standard deviation, s_g, using Equations (5) can pose problems when the observed concentration is close to zero and the corresponding model estimate is finite; the large logarithm of the ratio dominates the calculation. This can be avoided by equating m_g to the median of the ratio of the observed to predicted concentration ratio, and using the interquartile range of the ratios to estimate s_g.

We suggest that the correlation between the residual and the model estimate represents the potential for model improvement. We can quantify this correlation through the linear model:

$$\varepsilon = aC_p + b + \mu \, , \tag{6}$$

where a and b are constants, and μ is a random variable with a mean of zero, and is not correlated with C_p. The transformation in Equation (3), denoted by the symbol '^', has been dropped for convenience. We will refer to μ as the stochastic error. Then the least squares estimate of the slope a is given by

$$a = \frac{\left\langle C_p' \varepsilon' \right\rangle}{\sigma_p^2} \, , \tag{7}$$

where σ_p is the standard deviation of the model estimates. The value of a is a measure of the undesirable correlation between the error and C_p, and can be presumably reduced by improving the model. Equation (6) allows us to express the standard deviation of the error as a sum of two components:

$$\sigma_\varepsilon^2 = \left(a\sigma_p\right)^2 + \sigma_\mu^2 \, . \tag{8}$$

The first term on the right hand side is correlated with the model estimate, and can be reduced in principle. The second term is the inherent error associated with information not included in the model inputs. The concepts discussed here are illustrated through an example.

4. APPLICATION OF MODEL PERFORMANCE STATISTICS

In the example considered here, the concentrations are transformed using logarithms as in Equation (4). Figure 1 shows the performance of two models in explaining daily maximum ozone concentrations observed in Riverside, California, in 1992.

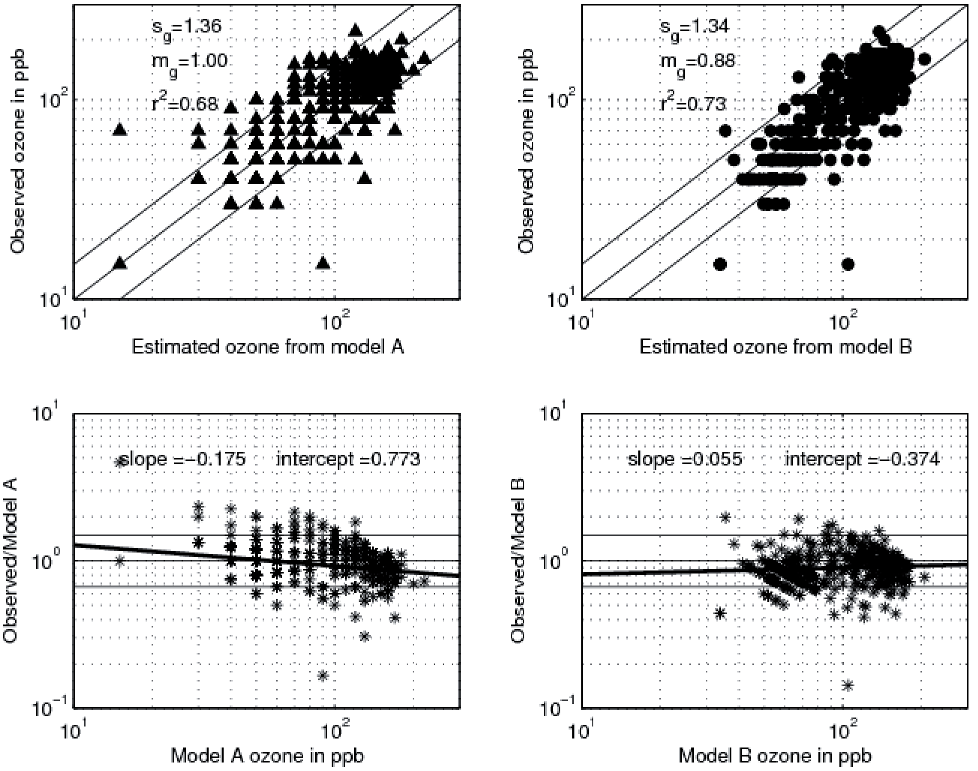

Figure 1. Performance of two models in explaining daily maximum ozone concentration measured in Riverside, California. Top panels show scatter plots and residual statistics. Bottom panels show behavior of residuals as a function of model estimates.

The top panels of the figure indicate the performance of the model in terms of the statistics, m_g and s_g and the coefficient of determination, r^2, between the logarithms of the observed and estimated ozone concentrations. If we assume that these error statistics are independent of the model estimate, we can use them to estimate the probability that the observation will exceed a certain value given the model estimate. Assuming that the observed concentrations are lognormally distributed about the model estimate, we can estimate the probability that the observed value corresponding to a model estimate exceeds 120 ppb, the US 1 hour standard. When the model estimate is 120 ppb, there is a 50% probability that the corresponding observation will exceed 120 ppb. The uncertainty in estimating an observed value increases with the value of the residual standard deviation, s_g. When $s_g=1.2$, the probability that the observed value exceeds 120 ppb is close to zero when the model estimate is 50 ppb. On the other hand, when s_g=2.5 and the model estimate is 50 ppb, the probability that the corresponding observation exceeds 120 ppb is close to 20%. When the model estimate is 150 ppb and s_g=1.2, the probability that the corresponding observation exceeds 120 ppb is 90%. For a model with s_g=2.5, this probability is only 65%. This exercise shows that there is no reason to classify a model as "good" or "bad" once we specify m_g and s_g for the model. These statistics provide information on the degree of uncertainty in the relationship between model estimate and corresponding observation.

The bottom panels of Figure 1 show the residual, $\varepsilon=ln(C_o/C_p)$, plotted as a function of C_p. The slope of the trend lines shown in the bottom panels of Figure 1 is a corresponding to Equation (5). We see that the slope, a, for model A is -0.18, which is about three times that of model B (0.06). The intercept, b, is a measure of the systematic error in the model, and is related to variables not included in the model inputs. It can be arbitrarily subtracted from the model, but one needs justification for doing so. We see that in absence of correlation between error and model estimate, the model performance statistic, m_g, would become $exp(b)$. The modified m_g would be $exp(-0.175)=0.84$ for model A and $exp(0.055)=1.06$ for model B, which are more representative of the performance of the models than the raw statistics would indicate. The next section describes a graphical depiction of the relationship between observations, model estimates, and model errors.

5. DISPLAYING MODEL PERFORMANCE GRAPHICALLY

Taylor (2001) proposed a diagram to display model performance. It is based on the relationship:

$$\varepsilon' = C_o' - C_p' \ , \tag{9}$$

where the prime refers to deviation from the mean and the model estimate and observation have been transformed appropriately. Squaring both sides of the lower equation and averaging results in

$$\sigma_\varepsilon^2 = \sigma_p^2 + \sigma_o^2 - 2r\sigma_p\sigma_o \ , \tag{10}$$

where the correlation coefficient, *r*, is given by

$$r = \frac{\left\langle C_p' C_o' \right\rangle}{\sigma_p \sigma_o} \equiv cos\,\theta \ . \tag{11}$$

Using the cosine rule, Taylor (2001) showed that the relationships between the variances in Equation (10) can be represented with the diagram of Figure 3.

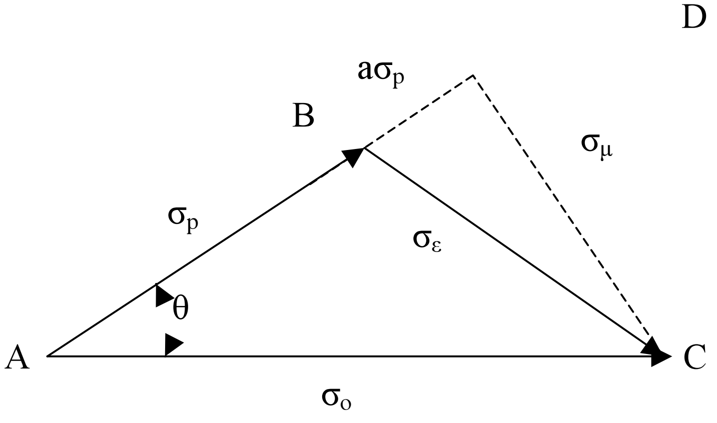

Figure 2. Geometric representation of Equation (9) adapted from Taylor (2001).

Taylor (2001) proposed comparing the relative performance of models by plotting the point B corresponding to each model in a diagram, where all lengths in the diagram are normalized by the standard deviation of the observations. The distance BC, represents model error, and the angle, θ, measures the correlation between model estimates and observations. These parameters can be compared to determine relative performance. We can add information on the behavior of the model error to the Taylor diagram by interpreting the diagram as a vector relationship. The component of the error vector, σ_ε, along the model estimate standard deviation is BD given by

$$BD = \sigma_\varepsilon \, cos\,C\hat{B}D = \sigma_\varepsilon \frac{\left\langle \varepsilon' C_p' \right\rangle}{\sigma_\varepsilon \sigma_p} = a\sigma_p \ , \tag{12}$$

which is consistent with Equation (8). The magnitude of BD relative to σ_o is a measure of the adequacy of the model. If all the lengths in the figure are normalized by σ_o, so that AC is unity, the length AD represents the correlation coefficient between observations and model estimates. The length CD, which is σ_μ, represents the component of the model error that is unrelated to model inputs.

Figure 3 uses the diagram of Figure 2 to compare the performance of two models used to estimate surface heat flux during unstable conditions. One is based on free convection theory, and other is an improvement suggested by Tillman, J.E. (1972). Now σ_p lies along the x-axis, and the radius of the arc is σ_o. The x-coordinate of each point is the correlation coefficient, and the y-coordinate is σ_μ, the inherent component of model error. The vector joining the point to the end of the σ_p vector is the total error, σ_ε. All lengths have been normalized by the standard deviation of the observations.

The free convection model has a correlation coefficient of 0.55, while that of Tillman's model is 0.67. The free convection model has an inherent error of 0.84, while Tillman's model has an error of 0.75. By these measures, Tillman's model is better than the free convection model. Note that the diagram provides additional information that can be used to improve the models. The potential for improvement of each model is given by the projection of the error vector on the x-axis. This projection is 0.3 for the free convection model, while it is 0.05 for Tillman's model. This suggests that the free convection model can be improved through reformulation of the model, while Tillman's model can be improved primarily through expansion of model inputs. This conclusion illustrates the value of an explicit model for the residual between model estimate and observation.

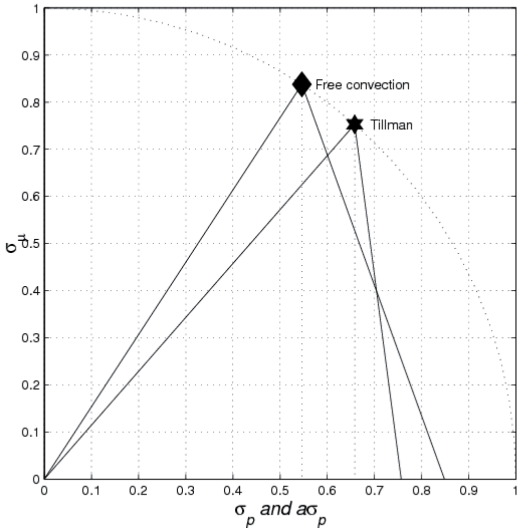

Figure 3. Comparison of the performance of three models using the diagram shown in Figure 2. The y-co-ordinate of each model point refers to, σ_μ , the component of the model error that is unrelated to model inputs. The x-axis plots the correlation coefficient, r, the standard deviation of the model estimates, σ_p, and the component of the model error, $a\sigma_p$, that is correlated with the model inputs.

6. SUMMARY

This note makes the case for formulating a statistical model to describe the behavior of the residuals between model estimate and corresponding observations. The distribution of residuals can be conveniently characterized in terms of the geometric mean, m_g, and the geometric standard deviation, s_g, of the ratio of observations to model estimates. These parameters can be used to estimate the probability density function (pdf) of observations corresponding to a model estimate. We show how this pdf can be used to answer questions of practical and perhaps regulatory relevance.

We then propose a linear model for the residual that allows separation of the model error into two components: one correlated with the model estimate, and an inherent component related to variables that are not part of the model input set. We show how this model of error can be depicted in a diagram that is similar to one proposed by Taylor, A. (2001).

REFERENCES

Csanady, G. T., 1973: Turbulent Diffusion in the Environment. D. Reidel Publishing Co.

Chang, J. C., S. R. Hanna, 2004: Air quality model performance evaluation. Meteorol. *Atmos. Phys.*, **87**, 167-196.

Taylor, A., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, **106**, 7183-7192.

Tillman, J.E., 1972: The indirect determination of stability, heat and momentum fluxes in the atmospheric boundary layer from simple scalar variables during dry unstable conditions, *Journal of Applied Meteorology,* **11**, 783–792.

Venkatram, A., 1982: A framework for evaluating air quality models. *Boundary-Layer Meteorology*, **24,** 371-385.

Weil, J.C., R I. Sykes, and A. Venkatram, 1992: Evaluating air quality models: Review and outlook. *J. Applied Meteorol.,* 1121-1145.