Some Comments on the Evaluation of Model Performance

Cort J. Willmott

Center for Climatic Research
Department of Geography
University of Delaware
Newark, Del. 19711

Abstract

Quantitative approaches to the evaluation of model performance were recently examined by Fox (1981). His recommendations are briefly reviewed and a revised set of performance statistics is proposed. It is suggested that the correlation between model-predicted and observed data, commonly described by Pearson's product-moment correlation coefficient, is an insufficient and often misleading measure of accuracy. A complement of difference and summary univariate indices is presented as the nucleus of a more informative, albeit fundamentally descriptive, approach to model evaluation. Two models that estimate monthly evapotranspiration are comparatively evaluated in order to illustrate how the recommended method(s) can be applied.

1. Introduction

It was a pleasure to read Fox's (1981) article on judging air quality model performance, as it represents one of a very few papers calling for the establishment of a consistent and rational set of procedures that should be used to evaluate model performance. Perhaps the only serious shortcoming of the paper is that its topical frame of reference is restricted to air quality models when many of the other fields within the atmospheric sciences also could benefit from such a discourse. Not only is there a paucity of literature on the general topic of model evaluation, but ambiguities can provide substantial frustration to those engaged in testing a model, comparing alternative formulations, or selecting the "best" model from the literature. With modeling becoming perhaps the major thrust within the atmospheric sciences, it is increasingly important for the discussion of model evaluation to be expanded in order that a generally accepted cadre of complementary methods and measures may soon emerge.

My purpose here, subsequently, is 1) to add to Fox's comments in order that they may have wider applicability and 2) to provide an alternative point of view—although I am in agreement with the main tenets of Fox's proposal. Consistent with the most frequently encountered model evaluation problems as well as with Fox, these comments are directed at the numerical comparison of observed and model-predicted variables where the variables are 1-dimensional (i.e., $N \times 1$), and their elements are scalar quantities. Many of the concepts, however, can be extended to the comparison of observed and model-predicted scalar and vector fields. My discussion focuses on the general utility and information content of the correlation and "difference" measures, although other topics, such as hypothesis testing and graphics, are briefly examined. Since there are far too many types of models within the atmospheric sciences to adequately discuss their "scientific" evaluation within a single paper, this discussion emphasizes "operational" evaluation, even though the scientific merit of any model—particularly one constructed for "explanatory" purposes (Mather *et al.*, 1980)—can be extremely important.

2. Correlation measures

Computing a quantitative index of association, covariation or correlation between an observed (O) and model-predicted (P) variate can take a variety of forms although, owing in some measure to historical inertia, Pearson's product-moment correlation coefficient (r) is chosen almost exclusively for the task. Sometimes its square, the coefficient of determination (r^2), is reported, but it provides little additional information. The proportion of the "variance explained" by P and embodied in r^2, however, offers a slightly more intuitively satisfying measure of model performance than mere correlation. At the same time, it is common to find the statistical significance of r or r^2 presented in order to corroborate interpretations of the correlation coefficient(s).

My most serious difference of opinion with Fox is related to the question of whether or not r or r^2 should be used at all. The main problem is that the magnitudes of r and r^2 are not consistently related to the accuracy of prediction, i.e., where accuracy is defined as the degree to which model-predicted observations approach the magnitudes of their observed counterparts. Willmott (1981), for instance, demonstrates that correlations between very dissimilar model-predicted variables and O can easily approach 1.0, while a number of recent comparisons of solar irradiance models illustrate that r and r^2 are insufficient to make meaningful distinctions between models (Powell, 1980; Davies, 1981; MacLaren Limited *et al.*, 1980). In another study, Willmott and Wicks (1980) observed that "high" or statistically significant values of r and r^2 may in fact be misleading, as they are often unrelated to the sizes of the differences between O and P. It is also quite possible for "small" differences between O and P to occur with low or even negative values of r. Since the relationships between r and r^2 and model performance are not well-defined, and not consistent, r or r^2 should not be part of an array of model performance measures.

0003-0007/82/111309-05$05.25

It is additionally inappropriate to report that such measures are statistically significant. Not only can this be a problem because the magnitude of r and its associated significance level are not necessarily related to the accuracy with which the model predicts O, but also because O and P (especially P) rarely conform to the assumptions that are prerequisite to the appropriate application of inferential statistics. Very few climatological papers, for instance, even test for or report the degree to which assumptions have been satisfied, even though such information is necessary for a meaningful interpretation of "significance." It is also rare to find justification for the selected level of type I error, let alone the indirectly chosen type II error. It can be concluded that neither r, r^2, or tests of their statistical significance have any real practical value in the evaluation of model performance.

3. Difference measures

Fox (1981) recommends, in essence, that four types of difference measures should be calculated and reported. Bias can be described by the mean bias error (MBE), while the variability of (**P—O**) about MBE (s_d^2) is merely the variance of the distribution of differences. Average difference can be alternatively described by the root mean square error (RMSE), its square—the mean square error (MSE), or the mean absolute error (MAE). Following Fox, these indices take the form

$$\text{MBE} = N^{-1} \sum_{i=1}^{N} (P_i - O_i) \tag{1}$$

$$s_d^2 = (N-1)^{-1} \sum_{i=1}^{N} (P_i - O_i - \text{MBE})^2 \tag{2}$$

$$\text{RMSE} = [N^{-1} \sum_{i=1}^{N} (P_i - O_i)^2]^{0.5} \tag{3}$$

and

$$\text{MAE} = N^{-1} \sum_{i=1}^{N} |P_i - O_i|, \tag{4}$$

where N is the number of cases. Clearly RMSE and MAE are among the "best" overall measures of model performance, as they summarize the mean difference in the units of O and P. As Fox mentions, MAE is less sensitive to extreme values than RMSE, and it should be added that MAE is intuitively more appealing, since it avoids the physically artificial exponentiation that is an artifact of the statistical-mathematical reasoning from which RMSE comes. On the other hand, MSE and RMSE are generally amenable to more in-depth mathematical or statistical analyses than MAE. Nonetheless, since MAE and RMSE are similar measures, it is appropriate, in many cases, to report either or both indices. The first two moments of the distribution of differences (MBE and s_d^2), however, do not provide enough diagnostic value to justify their inclusion, over other measures, in an array of model evaluation measures.

The MBE is merely the difference between the mean of the model-predicted variable ($\bar{P}$) and the observed variable ($\bar{O}$). Common sense then suggests researchers might alternatively report $\bar{P}$ and $\bar{O}$, since more people are familiar with these indices, and they contain a little more information than MBE by itself. When MSE, in addition to $\bar{P}$ and $\bar{O}$, is known, s_d^2 also adds little "new" information, as it is a simple function of the above terms, i.e.,

$$s_d^2 \simeq \text{MSE} - (\bar{P} - \bar{O})^2 \tag{5}$$

when N becomes large. Perhaps s_d^2 could be interpreted as the average "noise" or unbiased difference, but other measures, e.g. the "unsystematic" difference (described in an upcoming paragraph), are more representative in their summary of the noise level. It seems that the third or fourth moment might supply, in this case, more useful information than either the first (MBE) or the second (s_d^2) moment of the distribution of differences. Of all the indices proposed by Fox, only MAE or RMSE and N really need to be computed and reported when they are accompanied by certain additional measures.

Summarizing points made in another paper (Willmott, 1981), investigators should, as a minimum, compute, interpret, and report the following "summary measures": $\bar{P}$, $\bar{O}$, the standard deviation of the predicted variable (s_p), the standard deviation of the observed variable (s_o), the intercept (a) and slope (b) of the least-squares regression, $\hat{P}_i = a + bO_i$. These measures have the advantages of being well known and well understood, and a variety of measures can be easily computed from these for special purposes. Not only do $\bar{P}$, s_p, $\bar{O}$, and s_o generally describe the two variates, but a and b together are much more illuminating than r and r^2, with respect to the nature of the linear covariance between P and O. As mentioned above, these summary measures should accompany the appropriate difference measures.

Difference measures are all derived from the fundamental quantity $(P_i - O_i)$, although each measure is scaled in a different way in order to describe particular features of the magnitudes of the differences. The measures MAE and RMSE give estimates of the average error, but neither measure provides information about 1) the relative size of the average difference or 2) the nature (type) of the differences comprising MAE or RMSE. Relative difference measures, such as RMSE/$\bar{O}$ or MSE/s_o^2, occasionally appear in the literature, but the general utility of such indices is questionable because they are unbounded, and unstable when $\bar{O}$, s_o, and/or N becomes small (near zero). Working from MSE, Willmott (1981, 1982) and Willmott and Wicks (1980) alternatively proposed and used an "index of agreement" (d) of the form

$$d = 1 - \left[\sum_{i=1}^{N} (P_i - O_i)^2 / \sum_{i=1}^{N} (|P_i'| + |O_i'|)^2 \right], \quad 0 \leqslant d \leqslant 1 \tag{6}$$

where $P_i' = P_i - \bar{O}$ and $O_i' = O_i - \bar{O}$. The index (d) is intended to be a descriptive measure, and it is both a relative and bounded measure which can be widely applied in order

to make cross-comparisons between models. Because a model ought to "explain" most of the major trends or patterns present in O, it also is important to know how much of RMSE is "systematic" in nature and what portion is "unsystematic." That is, with respect to a "good" model, the systematic difference should approach zero while the unsystematic difference approaches RMSE. In order to make quantitative estimates of these types of error, Willmott (1981) proposed that the systematic error can be described by

$$\mathrm{MSE_s} = N^{-1} \sum_{i=1}^{N} (\hat{P}_i - O_i)^2 \tag{7}$$

while the unsystematic error can take the form

$$\mathrm{MSE_u} = N^{-1} \sum_{i=1}^{N} (P_i - \hat{P}_i)^2. \tag{8}$$

As the system is conservative,

$$\mathrm{MSE} = \mathrm{MSE_s} + \mathrm{MSE_u}. \tag{9}$$

(Computational forms are given in Willmott (1981)). Eqs. (7), (8) and (9) are useful in that the proportion of MSE that arises from systematic errors, presumably contained in the model, is described by ($\mathrm{MSE_s}$/MSE) and the unsystematic proportion is ($\mathrm{MSE_u}$/MSE) or [1 − ($\mathrm{MSE_s}$/MSE)]. Moreover, these differences can be interpreted, in the units of P and O, by taking the square roots of $\mathrm{MSE_s}$ and $\mathrm{MSE_u}$, i.e., $\mathrm{RMSE_s}$ and $\mathrm{RMSE_u}$, respectively. It is suggested, subsequently, that $\mathrm{RMSE_s}$, $\mathrm{RMSE_u}$, and d should be computed, interpreted, and reported in addition to RMSE or MAE and the above-mentioned summary measures.

Fox also indicates that confidence interval estimates about MBE, for instance, and hypothesis tests of MBE, $s_{\bar{d}}^2$ and MSE, can be helpful, although he concedes that ". . . statistical significance of model performance measures will not be established easily." As alluded to during my discussion of r and r^2, experience suggests that confidence bands and tests of statistical significance are not nearly as illuminating as an informed scientific evaluation of the summary and difference measures. Scientific evaluation can be further enhanced by the examination of data-display graphics and sensitivity analyses, as well as by comparisons with other models. Without reiterating arguments that have been made a number of times before, suffice it to say that my mistrust of statistical significance and hypothesis testing in the area of model evaluation arises from numerous observations that "real" significance and statistical significance do not exhibit a close correspondence. It is appropriate perhaps to report the magnitude, probability, and degrees of freedom associated with an F, t, χ^2, or Kolmogorov-Smirnov statistic, but given how little is generally known or reported about the parent distributions and sampling biases contained in P and O, for instance, it is misleading to argue, on statistical grounds, that MSE or any other measure is "significant." Evaluations of the relationships between P and O should alternatively be based upon the axioms of mathematics, knowledge of the sensitivity of the summary and difference measures to patterns and anomalies within P and O, understanding of the model and the processes it attempts to paraphrase, and the reliability of the model-input and observed (test) data and the computational scheme employed.

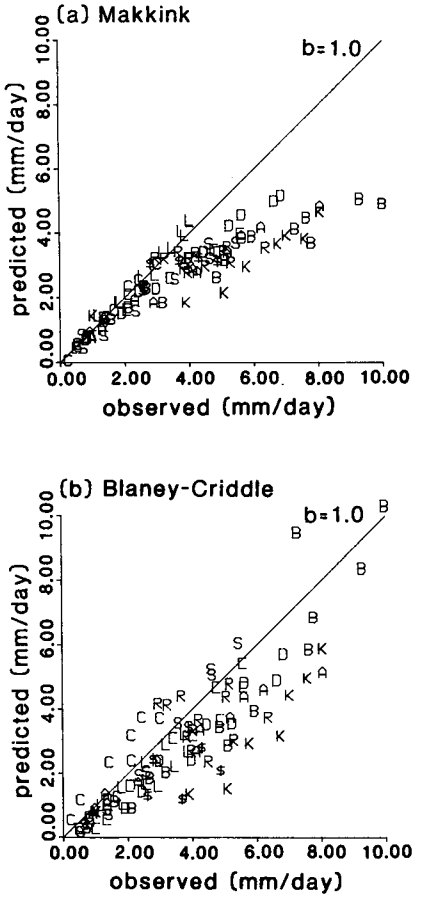

FIG. 1. Scatterplots of monthly lysimeter-derived (observed) evapotranspiration at 10 stations versus two model-predicted variables for those same months and station locations. The model-predicted variables were generated by the models of a) Makkink and b) Blaney-Criddle, and a unique symbol is used to identify observations associated with a particular station. The stations and their corresponding plot symbols (given in parentheses) are Aspendale, Australia (A); Brawley, Calif. (B); Copenhagen, Denmark (C); Coshocton, Ohio (C); Davis, Calif. (D); Kimberly, Idaho (K); Lompoc, Calif. (L); Ruzizi Valley, Zaire (R); Seabrook, N.J. (S); and South Park, Colo. ($). The perfect prediction line ($P_i = O_i$) also is plotted and labelled ($b = 1.0$) for reference.

4. An evaluation of two evapotranspiration models

Observed and model-predicted data that were interpolated from graphs given by Jensen (1973) are used to evaluate two climatic models that estimate monthly evapotranspiration from weather, irradiance, and/or site data in order to illustrate the relative utility of the above-discussed evaluation measures. Forms of the Makkink and Blaney-Criddle models that were implemented by Jensen are presented because they exemplify the disagreement that can occur between the

TABLE 1. Quantitative measures of evapotranspiration model performance.*

	$\bar{O}$	$\bar{P}$	s_o	s_p	N	a	b	MAE	RMSE	$RMSE_s$	$RMSE_u$	d	r^2
Makkink	3.65	2.69	2.17	1.25	104	0.80	0.52	1.03	1.52	1.42	0.55	0.82	0.80
Blaney-Criddle	3.65	2.90	2.17	1.95	104	0.01	0.79	1.00	1.28	0.88	0.94	0.90	0.77

*The terms N, b, d, and r^2 are dimensionless, while the remaining terms have the units mm d^{-1}.

correlation and difference measures. Virtually no attempt is made to examine the physical-mathematical bases of these models or the quality of O (it is assumed to be error-free) since the emphasis is on the evaluation measures. If the reader desires a more in-depth discussion of these models and data, it is suggested that Jensen (1973) be consulted.

Since data-display graphics (also recommended by Fox) can be extremely helpful in identifying the pattern of the differences between P and O as well as extreme cases, the graphics should always accompany the quantitative indices. Scatterplots, in particular, can well represent the relationships between P and O. Highly dissimilar covariance patterns become apparent between Jensen's observed variable (O^J) and the model-predicted variates derived from the algorithms of Makkink (P^M) and Blaney-Criddle (P^B), when they are displayed in scatterplot format—for example (Fig. 1). It should be noted that the superscripts B, M, and J are used to identify those variables and indices associated with Blaney-Criddle's model, Makkink's method, and Jensen's observed data, respectively. From the plots alone, it appears that Makkink's model is less variable than Blaney-Criddle's method but, at the same time, P^M systematically underestimates O^J—to a much greater degree than P^B. When O_i^J is less than 4.0 mm d^{-1}, Makkink's model seems preferable while, over the upper range of O^J, the Blaney-Criddle approach appears more accurate. These brief interpretations, which are intended to illustrate diagnoses that can be made from graphs, are supported by most of the quantitative measures.

Examination of the summary position and scale parameters ($\bar{P}$ and s_p) indicates that $\bar{P}^B$, $\bar{P}^M$, s_p^B and s_p^M all underestimate the corresponding observed parameters $\bar{O}^J$ and s_o^J (Table 1). Consistent with the graphic interpretation, $\bar{P}^M$ is in error by about 26% while $\bar{P}^B$ posts a somewhat smaller difference of nearly 21%. At the same time, Blaney-Criddle's method does substantially better than Makkink's model at predicting the variability contained in O^J. The regression parameters (a and b) suggest similar, systematic (linear) underpredictions, although they should not be separately interpreted as additive and proportional differences—they are mathematically dependent on each other. A more comprehensive evaluation can be made from the difference indices.

With respect to MAE, no meaningful distinction can be made between P^M and P^B, but RMSE suggests that P^B is 0.24 mm d^{-1} closer to O^J than P^M—on the average (Table 1). Once again, the difference between MAE and RMSE results from the weighting of each $(P_i - O_i)$ by its square which tends to inflate RMSE, particularly when extreme values are present. The root mean square error, therefore, can be generally regarded as a high estimate of actual average error or MAE, except when a significant number of the $|P_i - O_i|$ are less than 1.0. In this example, the credibility of the difference between $RMSE^B$ and $RMSE^M$ is enhanced because it is consistent with trends described by $\bar{P}$, a, and b. The index of agreement more precisely suggests that P^B is about 8% more accurate than P^M, while r^2 erroneously indicates that P^M is the more accurate of the two. Disagreement between r^2 and the other measures characterizes its inadequacy when used as a measure of accuracy. Still, other useful information is contained in $RMSE_s$ and $RMSE_u$.

Since differences described by $RMSE_s$ can be described by a linear function, they should be relatively easy to dampen by a new parameterization of the model. In other words, without making significant changes in a model's structure, it should be possible to substantially reduce $RMSE_s$, which implies that $RMSE_u$ can be interpreted as a measure of potential accuracy. If this statistical point of view is reasonable and no physical-mathematical or other scientific information is considered, it can be concluded that Makkink's model is potentially more accurate. Regardless of whether or not accuracy or potential accuracy is evaluated, it is clear that no single index can adequately describe model performance and, therefore, researchers should report an array of complementary measures as suggested by Fox (1981), Willmott (1981), and others. This necessarily brief sample evaluation points to only a very few of the possible inferences that can be drawn from an eclectic evaluation of model performance.

5. Summary

A recent paper by Fox (1981) on judging air quality model performance has provided the impetus and basis for a more general discussion of documenting and evaluating model performance in the atmospheric sciences. It has been argued that the commonly used correlation measures such as r and r^2 and tests of statistical significance in general are often inappropriate or misleading when used to compare model-predicted (P) and observed (O) variables. Difference measures, however, seem to contain appropriate and insightful information. It is recommended that researchers compute and report the root mean square error or the mean absolute error as well as their systematic and unsystematic proportions or magnitudes, and the average relative error represented by the index of agreement. The interpretation of these measures should be descriptive and based on scientific grounds, not on the basis of the measures' statistical significance. Simple summary statistics and graphics also can help illuminate the relative ability of a model to predict accurately. An exemplary evaluation of two models that estimate monthly evapotranspiration is presented, in order to illustrate the recommended methodology.

References

Davies, J. A., 1981: *Models for Estimating Incoming Solar Irradiance.* Canadian Climate Centre, Downsview, Ontario, 101 pp. (Unpublished manuscript.)

Fox, D. G., 1981: Judging air quality model performance: A summary of the AMS Workshop on Dispersion Model Performance. *Bull. Am. Meteorol. Soc.*, **62**, 599–609.

Jensen, M. E. (ed.), 1973: *Consumptive Use of Water and Irrigation Water Requirements.* American Society of Civil Engineers, New York, N.Y., 215 pp.

MacLaren Limited, J. F., Hooper and Angus Associates Limited, J. E. Hay, and J. A. Davies, 1980: *Define, Develop and Establish a Merged Solar and Meteorological Computer Data Base.* Canadian Climate Centre, Downsview, Ontario, 179 pp. (Unpublished manuscript.)

Mather, J. R., R. T. Field, L. S. Kalkstein, and C. J. Willmott, 1980: Climatology: The challenge for the eighties. *Prof. Geogr.*, **32**, 285–292.

Powell, G. L., 1980: *A Comparative Evaluation of Hourly Solar Global Irradiance Models.* Ph.D. Dissertation, Arizona State University, Tempe, Ariz., 240 pp.

Willmott, C. J., 1981: On the validation of models. *Phys. Geogr.*, **2**, 184–194.

——, 1982: On the climatic optimization of the tilt and azimuth of flat-plate solar collectors. *Solar Energy*, **28**, 205–216.

——, and D. E. Wicks, 1980: An empirical method for the spatial interpolation of monthly precipitation within California. *Phys. Geogr.*, **1**, 59–73. •

---

announcements (*continued from page 1308*)

First Year Ozone Results Available for Nimbus-7 SBUV/TOMS

Nimbus-7 Solar Backscattered Ultra-violet/Total Ozone Mapping Spectrometer (SBUV/TOMS) total ozone and ozone profile data are available for November 1978 to November 1979. The quality of the archived data is summarized in a validation statement provided by the Nimbus Experiment Team. A list of specific products that are available follows and instructions for obtaining the data are provided.

Total ozone and ozone vertical profile results for the SBUV/TOMS operation from November 1978 to November 1979 are available. The algorithms used have been thoroughly tested, the instrument performance examined in detail, and the ozone results compared with Dobson, Umkehr, balloon, and rocket observations. The accuracy and precision of the satellite ozone data are good to at least within the ability of the ground truth to check and are self-consistent to within the specifications of the instrument.

The primary input to the ozone retrieval algorithms is the ratio of the backscattered radiance to the incidence solar radiance. Both radiance and irradiance are measured separately by the SBUV and TOMS instruments. Accuracy in the determination of this ratio is better than 0.5% for SBUV and 1.0% for TOMS. Prelaunch calibration uncertainties affect the absolute accuracy of these measurements; the magnitude of these uncertainties is being assessed. During the first year of instrument operation in-flight diffuser degradation is less than 1.5% at 339.8 nm and less than 3.0% at 273.5 nm. No correction for this has been applied to the data. This in-flight degradation can introduce an apparent long term drift in an analysis of the data.

One of the major design improvements of the Nimbus 7 SBUV instrument over the BUV instrument on Nimbus 4 involved the employment of a system for the on-board subtraction of dark current. This has improved the instrument's performance to a point where no radiation induced signal has been observed in the South Atlantic anomaly to date.

Total ozone has been derived from both the SBUV and TOMS instruments. Analysis of the variance of comparisons between co-located TOMs and AD pair direct sun (OO code) Dobson observations shows that total ozone retrieval precision is better than 2% to within 10 degrees of the solar terminator. There are biases of −6.5% and −8.3% for TOMS and SBUV respectively when compared to the Dobson network. These biases are primarily due to inconsistencies in the ozone absorption coefficients used by the space and ground systems. If absorption coefficients available on a preliminary basis from the National Bureau of Standards were used for both SBUV/TOMS and the Dobson measurements, the biases would be less than 3%.

Vertical profiles of ozone have been derived from the SBUV step scan radiances at 273.5 nm and longer wavelengths using an optimum statistical inversion algorithm. Comparison with Umkehr measurements at Boulder and Arosa indicate that the precision of the derived layer ozone amounts is on the order of 5%.

The altitude range (for which the inferred SBUV profiles is determined primarily by the radiance measurements) depends on several factors including the solar zenith angle, the total ozone amount, and the shape of the ozone profile. This altitude range typically extends from 0/7 mb (50 km) down to the peak of the ozone density profile (20–40 mb) or 22–26 km. The derived layer ozone amounts below this region as produced by the optimum statistical inversion algorithm depend on the *a priori* statistical information about the correlation between these layers and the observed total ozone and upper level profile amounts.

Variations of UV solar flux associated with the rotation of active regions on the sun (27-day solar rotation period) have been observed with the continuous scan solar observations (160–400 nm). However, no significant solar flux variation was observed at the wavelengths used for the total ozone and vertical profile retrievals. Therefore, a long term smoothed solar flux was used in the processing with no 27 day period component. To the extent that there may have actually been a small 27 day period in the real solar flux this would introduce a small (less than 0.5%) artifact in the data.

Ground truth for the SBUV ozone profile consists of Umkehr profiles, ozone balloon sondes, several optical rocket sondes, and a chemiluminescent rocket sonde. This set of data

(*continued on page 1317*)