ON THE VALIDATION OF MODELS

Cort J. Willmott
Center for Climatic Research
Department of Geography
University of Delaware

Abstract: Traditional methods of evaluating geographic models by statistical comparisons between observed and simulated variates are criticized. In particular, it is suggested that the correlation coefficient (r), its square and tests of their statistical significance are inadequate for such purposes. The root mean squared error (RMSE) and related measures as well as a new index of agreement (d) are alternatively presented as superior indices for making such comparisons. Arguments are made for increasing the number of digital algorithms and data plots being published.

INTRODUCTION

Geographers, having finally begun to adopt a "model-based paradigm" (Haggett and Chorley, 1967), are now developing a myriad of algorithms to predict an almost infinite variety of events. Many of these models are based upon curve fitting procedures where a portion of the variability in a dependent variable is functionally linked to variabilities displayed in one or more independent variables. In other cases, such empirical functions are combined with other empirical functions or with deterministic or stochastic procedures to form more complex, semi-empirical or quasi-deterministic models.

Regardless of how one categorizes geographic models, one of the major goals of such quantitative systems is to predict an outcome, in space or time, given some input. The accuracy of the outcome, however measured, relative to its total cost is undoubtedly one of the most important measures of the model's worth. In order to make informed evaluations, however, predicted values must be compared with measured values in a meaningful, quantitative way.

Geographers have reported a variety of standardized coefficients of agreement or association as proof that their models accurately forecast empirical events. Most often, however, Pearson's Product-Moment Correlation Coefficient (r) is selected for the task since it describes colinearity between observed (O) and predicted (P) variates where the observations are interval or ratio in scale. Often in lieu of r, geographers report the coefficient of determination (r^2) which is perhaps an even better measure of the model's worth since r^2 describes the proportion of the total variance explained by the model. In order to assure the reader that the model is a good predictor of reality, a test for the statistical significance of the colinearity expressed in r or r^2 is usually reported. In fact, r, r^2 and tests of their significance have been accepted by geographers as *the* most reliable measures of a model's ability to estimate events. Often, however, the ability of a model to predict events is too elusive to be adequately encapsulated by these standardized coefficients of agreement or association (McCuen and Snyder, 1975)–and significance tests in general.

Owing to the problems associated with the above-mentioned standard measures and the ever pressing need to validate our models, an alternative set of indices may be calculated, evaluated and reported when comparisons between simulated and observed values are made. In particular, it is recommended that the following groups

Physical Geography, 1981, 2, 2, pp. 184-194

of descriptors be computed and reported:

1. Observed and predicted means ($\bar{O}$ and $\bar{P}$, respectively) and standard deviations (s_o and s_p, respectively);
2. Slope (b) and intercept (a) of a least-squares regression between the predicted (dependent variable) and observed (independent variable);
3. Systematic and unsystematic components of the root mean squared error ($RMSE_s$ and $RMSE_u$, respectively) as well as the total root mean squared error (RMSE);
4. Index of agreement (d).

Only the more essential diagnostic measures are listed since a wide variety of other measures, including the standard measures, can be easily computed from these for more special purposes. Preliminary points in favor of (1) publishing computer programs, (2) dropping the testing for and reporting of statistical significance, and (3) increasing the number of published data plots that describe the relationships between observed and simulated variables will precede the discussion of the validation statistics. Arguments for calculating and reporting indices from the above four categories will then be presented. Although my comments reflect a personal viewpoint, there is increasing support in the statistical literature for many of these opinions (Freedman et al., 1978).

PUBLISHING COMPUTER PROGRAMS

Over the past few years far too few computer programs have been published, resulting in the development of numerous overlapping and/or redundant algorithms which have been created by researchers who are either ignorant of the work of others or who simply cannot afford the time necessary to turn pertinent journal articles into functioning programs. In fact, most journal articles do not contain all the information necessary to reproduce the digital model from which the article's results were generated. In such cases, the program itself *is* the model, not those few significant expressions which are aired in the journals because of manuscript length and page cost restrictions. As models become ever more lengthy and complex, it will become increasingly important to have access to source programs in order to serve the guidelines of science, i.e., verification and reproducibility. Extensive algorithms, of course, cannot be published in the major journals; however, they should appear, along with the necessary documentation, in the appropriate monograph or technical series.

TESTING FOR STATISTICAL SIGNIFICANCE

Statistical significance is a concept which should be viewed with a large degree of scepticism (Freedman et al., 1978). It is perhaps appropriate to test an agreement measure (e.g., r^2) and to report the value of a test statistic (e.g., F) as well as the observed significance level (α), but to draw a line between "significant" and "insignificant" levels of agreement is completely unjustified. In most cases, α corresponds to a continuous distribution on which there is *no* significant difference between $\alpha = 0.049$ and $\alpha = 0.051$, for instance. Such a distinction is particularly uncalled for when it is unknown to what extent the assumptions underlying the test have been violated, and the power of the test is as much a function of N, the test itself and the sample distributions as it is of the true relationships contained in the data. It is again suggested that the value of α and a test statistic may or may not be reported, at the researcher's discretion, but *no* extra credence should be attached to the model because r or r^2 is statistically significant. Modelers must simply make their interpretations about a model's predictive worth on the basis of their knowledge of the processes the model describes, similarity and error indices, the accuracy of the input and test data, and the numerical or computational scheme employed.

DATA PLOTS

Data plots ought to accompany any comparison between observed and simulated variables as such graphic aids lend visual credibility to quantitative comparisons as well as point to possible erroneous computations. Observed versus predicted scatterplots, in particular, can illuminate the relationships between the two variables although the publication frequency of such graphs, in even the most model-oriented of the geographic journals, is surprisingly low. Geographers are not alone in their underuse of statistical graphics and, as a result, data plotting has been a topic of renewed discussion in the statistical literature where a number of new data plotting methods are being explored (McGill et al., 1978). Data plots (e.g., scatterplots, box-plots, plots of residuals and histograms) represent the first level of comparative data description and analysis, whereas summary univariate measures comprise the second level.

SUMMARY UNIVARIATE STATISTICS

Observed and predicted means and standard deviations should always be computed and reported since they describe perhaps the two most important parameters associated with interval and ratio scale data—position and scale. At the same time, most geographers understand them and, subsequently, can use them in making simple comparisons between observed and predicted variables. One of the more important uses for means and standard deviations, however, is in the easy computation of a variety of more diagnostic or special purpose statistics. A number of these can be economically calculated when the statistics $\overline{O}$, $\overline{P}$, s_o, s_p and b are all that is known.

THE PARAMETERS a AND b

A model that exactly reproduces the magnitudes of the empirical observations would produce a scatterplot between P and O on which all points fell on a line that had a slope (b) of 1.0 and an intercept (a) of 0.0. Since models rarely, if ever, produce such good results, the parameters a and b can be estimated by some form of least-squares procedure. The debate is not yet resolved, however, concerning which type of least-squares procedure is most appropriate for computing the linear relationship between two variables (Jones, 1979; Mark and Peucker, 1978).

Computation of the best line can range from the common practice of regression P on O where the residuals are parallel to the P axis (e.g., Atwater and Ball, 1978), to regressing O on P where the residuals are parallel to the O axis (e.g., Satterlund, 1979), to regressing O and P where the residuals are computed orthogonally to the regression line. The best function most often lies somewhere between these, although its accurate selection is, in many cases, not feasible as adequate theory is not yet available (see Jones, 1979).

Systematic (linear) over- or under-predictions produce characteristic variations in a and b. These parameters should, therefore, be reported in addition to $\overline{P}$, $\overline{O}$, s_p and s_o in order that they may be interpreted and used in partitioning the major sources of error. They (a and b) are particularly useful in the economical computation of systematic and unsystematic components of the average or mean squared error—under the assumption that any nonlinear error is unsystematic. As this assumption may be unrealistic in a few situations, an appropriate transformation might be required to convert the systematic dependence of P on O to one which is fundamentally linear.

Under the propositions that P is dependent upon O and the dependence is linear, the problem can be stated as an ordinary least-squares fit. The former assumption is of particular importance as it implies that all of the error variance is contained within P and O is error free. Strictly speaking, O is rarely, if ever, error free although with good data the assumption is quite reasonable. The alternative is to make a priori assumptions about the underlying nature of the two error variances which seems even less appropriate, in this case, because little is generally known about the observed and predicted populations. If one disagrees with this position, a and b can be alternatively estimated by procedures described in Jones' (1979) review paper. Subsequent calculations, described here, should then be updated in response to the way a and b are computed and interpretations will change. Upon accepting these two assumptions and computing a and b, one is then able to decompose the average error into its major parts.

ERROR INDICES

Average error produced by a model is encapsulated in the mean squared error (MSE) or its square root, the root mean squared error (RMSE). Neither statistic, unfortunately, is commonly reported by geographers although there are exceptions (e.g., Hay, 1978; Willmott and Wicks, 1980). RMSE, however, is frequently used to evaluate models in the physical sciences such as meteorology and physical oceanography (e.g., Atwater and Ball, 1978). Calculated as

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

the root mean squared error is easy to interpret since it has the same metric as O and P. It is an important validation statistic in that it informs the modeler and reader about the actual size of the error produced by the model—unlike r or r^2 in which a large error may be masked by high values of s_o and s_p or a small error may appear significant owing to low magnitudes of s_o or s_p. Even though RMSE has a number of advantages over r and r^2, it does not illuminate the sources or types of error which can be of considerable help in refining a model. It is particularly convenient if you know how much of MSE is systematic and how much is unsystematic—again in the linear frame. Moreover, it can be helpful to know how much of the systematic error is additive and how much is proportional. As knowledge about the magnitudes of these errors can further enhance understanding of a model's predictive ability, a description of their computation, assuming a and b are known, follows.

Additive systematic error (MSE_a), resulting from a constant over- or under-prediction of the observed values, can be expressed as

$$MSE_a = a^2 \tag{2}$$

while proportional systematic error (MSE_p) may be written

$$MSE_p = (b-1)^2 \left[N^{-1} \sum_{i=1}^{N} O_i^2\right]. \tag{3}$$

The proportional and additive components of MSE are, unfortunately, not completely independent of one another, and their interdependence can be described by

$$MSE_I = 2a(b-1)\bar{O}. \tag{4}$$

In other words, the degree of interdependence between additive and proportional errors is a function of the covariation of the additive error (a) and the proportionality error (b - 1). Even though MSE_a and MSE_p are not independent, their magnitudes, or the magnitudes of their square roots $RMSE_a$ and $RMSE_p$, can help both the modeler and reader in discerning the impact that each type of systematic error has on a model's predictive ability. The relative magnitude of MSE_I, or its signed square root[1] ($RMSE_I$) can also be instructive.

Probably of greater diagnostic importance, however, is the overall or average systematic error (MSE_s) which is merely the sum of the sources of systematic error, i.e.,

$$MSE_s = MSE_a + MSE_p + MSE_I. \tag{5}$$

Once MSE_s and MSE have been computed, the equally important unsystematic error (MSE_u) becomes

$$MSE_u = MSE - MSE_s. \tag{6}$$

Since the calculations of MSE_s and MSE_u are relatively simple, once a, b and MSE have been derived, they are presented as the computational form. More illuminating forms of MSE_s and MSE_u, however, are the sums of squares equations:

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

and

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

where $\hat{P}_i$ is derived from $\hat{P}_i = a + bO_i$. Interpretations of MSE, MSE_s and MSE_u are again made easier by examining their square roots–RMSE, $RMSE_s$ and $RMSE_u$, respectively–all of which have the units of O and P.

The utility and information derived by calculating and reporting MSE, MSE_u, MSE_s, MSE_a, MSE_p, MSE_I or their square or signed square roots are difficult to access for any given modeling effort. At the very least, they specify the magnitude and direction (sign) of the major types of error that, hopefully, will reflect on the sources of error which, in turn, can be corrected. Since the interpretation of $RMSE_a$, $RMSE_p$ and $RMSE_I$ is somewhat confounded by their interdependence, however, it is suggested that they may or may not be reported at the researcher's discretion. Conversely, RMSE, $RMSE_s$ and $RMSE_u$ should always be computed and reported as they describe adequately the magnitudes of the major errors. The importance of interpreting the new measures ($RMSE_s$ and $RMSE_u$), in addition to RMSE, is illustrated in the following paragraph.

When $RMSE_s$ and $RMSE_u$ are not computed and RMSE is entirely systematic but acceptable, an investigator may be tempted to accept the model as is. This is done because RMSE is assumed to be unsystematic. In such cases, however, further refinement is called for as $RMSE_s$ should be minimized in order that the model is predicting at its maximum possible accuracy. On the other hand, if RMSE is all, or largely, composed of $RMSE_u$, perhaps the model is as good as it can be without major reworking. In order to demonstrate the general use of these error indices, their application in an ongoing modeling effort by the author will be briefly discussed.

Over the past 2 years, an energy budget algorithm which predicts daily total global radiation, among other variables, has been under development. The routines were inspired by Vowinckel and Orvig (1972) and later updated by Willmott et al. (1978). Refinement of the shortwave procedures is still under way and one of the model's purposes is to predict daily total global radiation from three-hourly surface synoptic information and a radiosonde. To illustrate the use of the above-described indices, two forms of the model are compared (Fig. 1).

Global radiation Model 1 and Model 2 are identical except that Model 1 contains an error which produces excessive amounts of water vapor in three levels of a simulated atmosphere. The result of this error is systematic underpredictions of global radiation. In Model 2, the error has been corrected. The simulated and observed variates (O is the same in both scatterplots) correspond to a year's record (1975) at Sterling, Virginia (Fig. 1).

From the plots alone, it is difficult to ascertain the degree and type of influence which the error exerted on the model's predictive ability. The error measures, however, unmistakably exhibit the magnitude of the inaccuracy as well as the degree to which the correction improved the predictive accuracy over the year of record (Table 1). Model 1, for instance, contains a very high MSE_s ($\cong$84% of MSE) and this was ameliorated to such a high degree by the correction that Model 2 contains an MSE_s of only about 42% of the new MSE. With respect to the sources of $RMSE_s$, a sharp reduction in $RMSE_p$ with an accompanying increase in $RMSE_a$ is indicated although an exact interpretation is unwarranted as $RMSE_I$ is relatively large. Nevertheless, the next task in the refinement of the model will be to locate and improve those equations or procedures which might produce the sizeable additive error. A spatial application of RMSE, $RMSE_s$ and $RMSE_u$ can be found in Willmott and Wicks (1980).

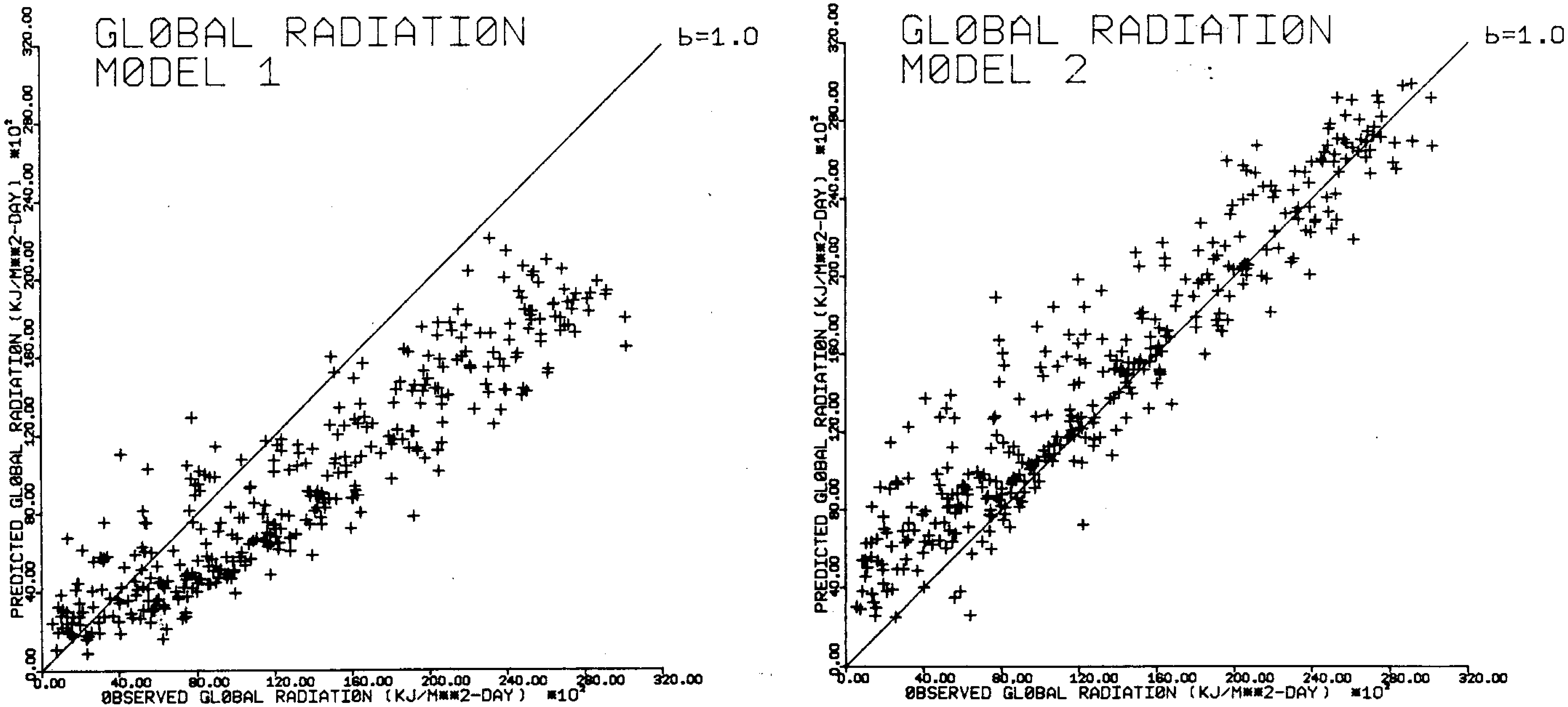

Fig. 1. Scatterplots of observed vs. predicted daily total global radiation (KJ $m^{-2}day^{-1}$) at Sterling, Virginia for 1975. The plots were produced by two slightly different versions of a previously published digital model (Willmott, Vowinckel and Orvig, 1978). N = 365.

Table 1. Global Radiation Summary Statistics[1]

	$\overline{O}$	$\overline{P}$	s_p	s_o	a	b
Model 1	13301.6	9536.2	7851.9	5385.0	1136.0	0.6315
Model 2	13301.6	14936.3	7851.9	71.03.0	3570.5	0.8545
	RMSE	$RMSE_u$	$RMSE_s$	$RMSE_a$	$RMSE_p$	$RMSE_I$
Model 1	5192.3	2100.1	4748.7	1136.0	5691.7	-3337.0
Model 2	3068.7	2332.2	1994.5	3570.5	2247.9	-3718.0
	MSE_u/MSE	MSE_s/MSE	MSE_a/MSE	MSE_p/MSE	MSE_I/MSE	d
Model 1	0.16	0.84	0.05	1.20	-0.41	0.86
Model 2	0.58	0.42	1.35	0.54	-1.46	0.96

[1]The dimensioned terms have the units KJ $m^{-2} day^{-1}$ and $N = 365$.

INDEX OF AGREEMENT

As previously stated, r and r^2 have been widely used by geographers to validate predictions made by their models. It is true that r and r^2 describe consistent proportional increases or decreases about the respective means of the two variates; however, they make too few distinctions between the type or magnitudes of possible covariations. To illustrate the problem, a hypothetical example is presented where both Model A and Model B are "perfect" predictors of the observed variable, as well as each other; that is, according to r and r^2 which both equal 1.0 in all cases (Fig. 2). The source of the inability of r and r^2 to distinguish between Model A and Model B is entirely systematic, in this case, and it stems from the statistics' inability to discern differences in proportionality and/or constant additive differences between the two variables.

In order to circumvent certain problems associated with r and r^2, an index of agreement (d) is presented. This new descriptive statistic reflects the degree to which the observed variate is accurately estimated by the simulated variate. The index of agreement is not a measure of correlation or association in the formal sense but rather a measure of the degree to which a model's predictions are error free. At the same time, d is a standardized measure in order that (1) it may be easily interpreted and (2) cross-comparisons of its magnitudes for a variety of models, regardless of units, can readily be made. It varies between 0.0 and 1.0 where a computed value of 1.0 indicates perfect agreement between the observed and predicted observations, and 0.0 connotes one of a variety of complete disagreements. Owing to its dimensionless nature, relationships described by d tend to complement the information contained in RMSE, $RMSE_s$ and $RMSE_u$.

Expressed as

$$d = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} [|P_i'| + |O_i'|]^2} \qquad (9)$$

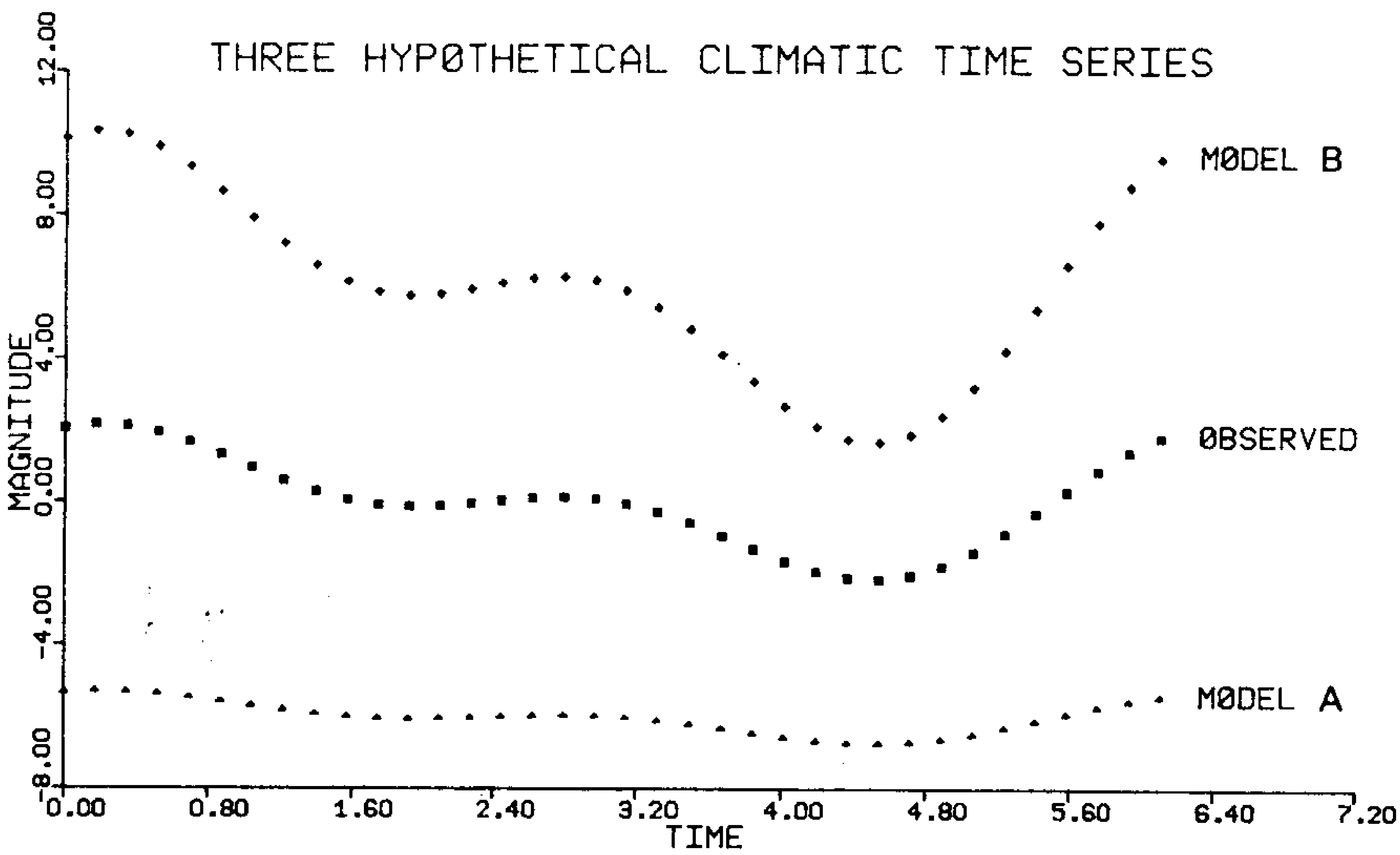

Fig. 2. Three hypothetical climatic time series in arbitrary units. Model A and Model B represent two dissimilar approaches to the simulation of the observed variable. N = 36.

where $\acute{P}_i = P_i - \bar{O}$ and $\acute{O}_i = O_i - \bar{O}$, d specifies the degree to which the observed deviations about $\bar{O}$ correspond, both in magnitude and sign, to the predicted deviations about $\bar{O}$. It is assumed that the portions of the magnitudes of P_i and O_i that are equivalent to $\bar{O}$ are not in error since $\bar{O}$ is considered to be error free. All the potential for error, in other words, is assumed to be contained in $\acute{P}_i$ and $\acute{O}_i$; or, more specifically, in the distances from each $\acute{P}_i$ to each $\acute{O}_i$. Therefore, the data set specific potential error (PE) is $\sum_{i=1}^{N} [|\acute{P}_i|+|\acute{O}_i|]^2$ or the sum of the squared absolute distances from P_i to $\bar{O}$ to O_i. The portion of this error left unexplained by the simulated variate is subsequently $\sum_{i=1}^{N} (\acute{P}_i - \acute{O}_i)^2$ or $\sum_{i=1}^{N} (P_i - O_i)^2$—the numerator of the main term in (9). When MSE has been computed previously, d is more economically obtained from

$$d = 1 - \frac{N \cdot MSE}{PE}. \tag{10}$$

Since d and the other statistics presented here are intended to be descriptive measures, the myriad of distributional assumptions associated with indices used for inferential purposes, save the axioms of algebra, are not necessary. It is further suggested that d's magnitude, again like the other measures presented here, be evaluated largely on the basis of knowledge about (1) the phenomena being studied, (2) data or measurement accuracy, (3) the simulation procedure or geographic model employed, and (4) the sensitivity of the descriptive statistics used to compare the observed and predicted variables.

Like other measures of agreement or association, the value of d lies in the nature of its response to the types of interrelationships that can exist in the data. For this reason, a variety of data sets were developed and evaluated by d as well as by the more commonly reported measures r and r^2. For brevity, the results of only two of those tests are described. The d statistic is also discussed in Willmott and Wicks (1980).

Unlike r and r^2, d is sensitive to differences between the observed and predicted means as well as to certain changes in proportionality. In the hypothetical example (Fig. 2), for instance, Model B is identified as being a slightly "better" estimator of the observed variate than Model A since its d value is 0.33 as opposed to 0.26 for the agreement between Model A and the observed variable. Because the distance between $\overline{O}$ and $\overline{P}$ is the same for both Model A and Model B, the implication is that Model B is a better covariate of the observed variate than Model A. Were the differences between $\overline{O}$ and $\overline{P}$ for the two models dissimilar, the values of the respective d's would reflect those errors as well. Again, it is interesting to note that the correlations between all combinations of Model A, Model B, and the observed variate are 1.0. In the "real world" example (Fig. 1; Table 1), d shows a dramatic improvement of Model 2 (d = 0.96) over Model 1 (d = 0.86) whereas r^2 posted a relatively small rise from 0.85 to 0.89. Visual examination of the time series' scatterplot (Fig. 1) tends to confirm the significance of the improvement.

Suffice it to say that in all cases the magnitudes of d were consistent with independent and intuitive evaluations made from graphs. As with r and r^2, the data set specific implications contained in d will only become intuitively meaningful after repeated use in a variety of problems.

SUMMARY AND CONCLUSIONS

Problems in the use of the correlation coefficient (r), its square and tests for their statistical significance to evaluate the predictive ability of models are described and evaluated. It is suggested that significance testing is inappropriate, in most situations, and the use of r or r^2 as a measure of accuracy is insufficient at best. Arguments are also made for publishing a greater number of digital models as well as data plots which visually describe the relationships between observed and simulated variables.

In lieu of the above statistics, it is proposed that the observed and predicted variates' respective means ($\overline{O}$, $\overline{P}$) and standard deviations (s_o, s_p); the intercept (a) and slope (b) of a least-squares linear regression between the variates; the errors described by the root mean squared error (RMSE), the systematic root mean squared error ($RMSE_s$) and the unsystematic root mean squared error ($RMSE_u$); as well as the index of agreement (d) should be computed and reported. These statistics, particularly RMSE, $RMSE_s$, $RMSE_u$ and d, are regarded as both adequate for and necessary to the interpretation and validation of predictions made by a model.

Acknowledgements: Support from the Computing Center and Center for Climatic Research at the University of Delaware is gratefully acknowledged. I also want to thank the following individuals who commented on earlier drafts of this paper: James E. Burt, Richard T. Field, James O. Huff, William D. Philpot, John N. Rayner, and John H. Schuenemeyer.

NOTE

[1]The sign function of a variable or function of a variable x is a standard FORTRAN procedure and it is written as Y = SIGN(f(x), x). The function f(x) is solved and the solution takes on the sign of x.

BIBLIOGRAPHY

Atwater, M. A. and Ball, J. T. (1978) A numerical solar radiation model based on standard meteorological observations. *Sol. Energy*, Vol. 21, 163-170.

Freedman, D., Pisani, R., and Purves, R. (1978) *Statistics.* New York: W. W. Norton & Co.

Haggett, P., and Chorley, R. J. (1967) Models, Paradigms and the New Geography. In R. J. Chorley and P. Haggett, eds., *Models in Geography.* London: Methuen & Co.

Hay, J. E. (1978) Measurement and modelling of shortwave radiation on inclined surfaces. In *Third Conference on Atmospheric Radiation* (Davis, Calif., June 1978), pp. 150-153. Boston: American Meteorological Society.

Jones, T. A. (1979) Fitting straight lines when both variables are subject to error. I. Maximum likelihood and least squares. *J. Int. Assc. Math. Geol.*, Vol. 11, 1-26.

Mark, D. M. and Peucker, T. K. (1978) Regression analysis and geographic model. *Can. Geogr.*, Vol. 22, 51-64.

McCuen, R. H. and Snyder, W. M. (1975) A proposed index for comparing hydrographs. *Wat. Resour. Res.*, Vol. 11, 1021-1024.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978) Variations of box plots. *Am. Statistn*, Vol. 32, 1-12.

Satterlund, D. R. (1979) An improved equation for estimating long-wave radiation from the atmosphere. *Wat. Resour. Res.*, Vol. 15, 1649-1650.

Vowinckel, E. and Orvig, S. (1972) EBBA—an energy budget programme. *Publs. Meteorol.*, No. 105, 1-50.

Willmott, C., Vowinckel, E., and Orvig, S. (1978) A modified digital model for predicting energy budgets in the United States. *Publs. Clim.*, Vol. 31, 1-170.

Willmott, C. J. and Wicks, D. E. (1980) An empirical method for the spatial interpolation of monthly precipitation within California. *Phys. Geogr.*, Vol. 1, 59-73.

ERRATA

In the Appendix of the article "An Empirical Method for the Spatial Interpolation of Monthly Precipitation within California" (*Physical Geography*, 1, 59-73) errors in equations (1), (2) and (3) were noted. The equations should read

$$d = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} \left[|P'_i| + |O'_i| \right]^2} \tag{1}$$

$$RMSE_s = \left[N^{-1} \sum_{i=1}^{N} (\hat{P}_i - O_i)^2 \right]^{0.5} \tag{2}$$

$$RMSE_u = \left[N^{-1} \sum_{i=1}^{N} (P_i - \hat{P}_i)^2 \right]^{0.5} \tag{3}$$