Concordance Correlation for Model Performance Assessment: An Example with Reference Evapotranspiration Observations

David W. Meek,* Terry A. Howell, and Claude J. Phene

ABSTRACT

The assessment procedures for agronomic model performance are often arbitrary and unhelpful. An omnibus analysis, the concordance correlation coefficient (r_c), is widely used in many other sciences. This work illustrates model assessment with two r_c measures accompanied with a mean-difference (MD) plot and a distribution comparison. Each r_c is an adjusted value of the usual Pearson correlation coefficient, r, assuming the exact relationship *observations = predictions*. The adjustments use a scale shift, u, and a location shift, v. Both of these measures also can indicate the similarity of the two variables' distributions; however, a formal test, the Kolmorgov-Smirnov D statistic, is used to statistically compare the distributions. Daily evapotranspiration data (ET_0) from a published study are compared with estimates from two possible weather observation based models. Although the first model has slightly lower r than the second (0.980 vs. 0.982), its predictions reasonably agree with observations by having comparatively small location and scale shifts [(u = 0.025, v = 1.10, D = 0.12 ($p \leq 0.86$)] and, consequently, a higher r_c (0.975 vs. 0.946). Results for the second model are comparatively unacceptable having larger scale and location shifts (u = 0.215, v = 1.19, D = 0.28 [$p \leq 0.04$]) with the bias ≠ 0 ($p \leq 0.05$) as clearly shown in the associated MD plot. Researchers should consider using r_c with an MD plot and distribution comparison in their model assessment toolkit because, together, they can provide a simple and sound probability based omnibus test as well as add useful insight.

Model performance assessment in agronomy and agroecology has employed many procedures in the past and continues to do so in the present. As recently reiterated by Tedeschi (2006), modelers in many fields, as well as statisticians, recommend the employment of multiple performance criteria tailored to each application rather than relying on a single performance measure. Relevantly, from this tailored application approach, Tedeschi (2006) provides formulae, discussion, and comparison of several available procedures.

With models becoming modular and increasingly complex, Fila et al. (2003) developed the public domain software library called IRENE_DLL (Integrated Resources for Evaluating Numerical Estimates- Dynamic Link Library). This library has most of the commonly used univariate and bivariate analyses including procedures called *difference based methods* and *association based methods*. Univariate analyses allow the modeler to consider and compare empirical and ideal distributional features of the observations and predictions. Difference based measures include methods such as the root mean square error, the mean absolute deviation or error, and the mean bias error (MBE) with a t test (see e.g., Fox, 1981). The MBE is often just called the *bias*. Association based methods include Pearson (r) or Spearman (r_s) correlation coefficients and simple linear least-squares regression analysis. In addition, there are empirical indices of agreement or efficiency like d (Willmott, 1981) and e (Nash-Suitcliffe efficiency statistic; Nash and Sutcliffe, 1970). Legates and McCabe (1999) proposed robust versions of d and e based on absolute value distances rather than squared distances measures. Other tools or updated versions of some of the IRENE tools are available and used in other disciplines, particularly medical research.

Using either a t test alone or a correlation coefficient alone can result in a misleading or inadequate assessment (Lin, 1989). In addition, many measures or indices, like d and e, are arbitrary, being ad hoc and not associated with significance tests and probability distributions. Willmott (1981, 1982, 1984), Willmott et al. (1985), and others argue against using significance tests. While many statisticians would agree with Willmott that assessment of model performance should make sense in terms of the underlying knowledge of the subject being modeled, they would still use the probability measures (Lin et al., 2002). Some statisticians, however, would recommend that no absolute significance level or confidence interval be preset, instead let the user decide whether or not to accept the results at the reported p levels. Moreover, in this spirit, know that an insightful omnibus procedure combining the essential

D.W. Meek, USDA-ARS, National Soil Tilth Lab., 2110 University Blvd., Ames, IA 50011; T.A. Howell, USDA-ARS Conservation and Production Research Lab., P.O. Drawer 10, Bushland, TX 79012; C.J. Phene (retired), USDA-ARS, SDI+, 13089 Wiregrass Lane, Clovis, CA 93619. Received 12 Nov. 2008. *Corresponding author (Dave.Meek@ars.usda.gov).

Published in Agron. J. 101:1012–1018 (2009).
doi:10.2134/agronj2008.0180x
Copyright © 2009 by the American Society of Agronomy, 677 South Segoe Road, Madison, WI 53711. All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Abbreviations: δ, Laio's location shift relative to scale; μ_x, prediction variable mean; μ_y, observation variable mean; A_p, Laio's bias correction factor; BV, bivariate; CIMIS, California Irrigation Management Information System; C_b, Lin's bias correction factor; d, Wilmott's index of agreement; D, Kolmorgov-Smirnov two-sample test statistic; e, Nash-Suitcliffe efficiency statistic; EDF, empirical distribution function; ET_0, daily evapotranspiration data; i, observation index; MBE, mean bias error (i.e., *bias*); MD, mean-difference; MSD, mean squared deviation; n, number of paired observations; r_c, concordance correlation coefficient; u, Lin's location shift relative to scale; x, prediction variable; x_i the ith prediction; y, observation variable; y_i the ith observation.

assessments of the difference and correlation analyses exists, is well-known, and is well-used in other scientific literature with 706 citations in the Web of Science as of February, 2009. This procedure is concordance correlation analysis. Lin (1989) first introduced the r_c as a bivariate analysis to assess agreement between paired measurements. The analysis, however, can be used for other similar assessments including comparing modeling predictions with independent observations (Lin et al., 2002). In addition Liao (2003) offers an improved concordance correlation procedure. To date, we found only two agronomic data comparison studies that use Lin's method, and neither is a modeling application (Joyce et al., 2001; McGinn et al., 2006). To promote the use of these procedures in agronomic modeling, this work presents a brief review of several procedures for comparing model performance. As an illustration, we assessed the performance of two reference evapotranspiration models using evapotranspiration data obtained from a lysimeter.

MATERIALS AND METHODS

Data Set and Models

Daily totals for hourly weighing lysimeter observations are taken from the 1985–1986 Phene et al. (1986) calibration study run at Five Points, CA. Corresponding daily totals of hourly modeled evapotranspiration were estimated with concurrently recorded hourly weather data collected from an adjacent meteorological station. Each series of the ET_0 and the resulting difference series have notable autocorrelations and so lack the desired degree of independence. The ET_0 (mm) from every other date for a total of 50 d are taken from one data set in the study. This selection reduces first-order serial correlation effects on the statistical analysis. The weighing lysimeter and ET_0 measurements are described in Howell et al. (1985). The weather station is described in Howell et al. (1984); the weather data set used in this study consists of direct hourly measurements of solar radiation, net radiation, air temperature, relative humidity, wind speed at 2 m elevation, wind direction, precipitation, and soil temperature at two depths. In Phene et al. (1986), the weather data were used to estimate evapotranspiration with several energy balance (EB) models (see e.g., Brutsaert, 1982) and then the modeling results were compared with the corresponding lysimeter measurements. The EB models are the sum of a radiation term and an advection term. The radiation term is mainly a function of the net radiation and humidity (cast as vapor pressure); the advection term is mainly a function and the wind speed and vapor pressure. For the purpose of this example, the daily lysimeter and weather observations are assumed to be valid having no bias or any other matters of concern.

The first model considered, Model-1, is the hourly modified Penman Equation with the empirical wind function of Pruitt and Doorenbos (1977) in the advection term as is reported in Phene et al. (1986). The Pruitt and Doorenbos (1977) hourly evapotranspiration model is the standard reference model for the California Irrigation Management Information System, CIMIS for brevity (Snyder et al., 1985, p. 52). The second model considered, Model-2, uses the same radiation term but the advection term is an iterative estimate from a physically based atmospheric stability routine (Eq. [10.20] on p. 218, Brutsaert, 1982). Here after it is called STAB for brevity.

Model Performance Measures

This example makes use of a multiple performance measures approach. Here the dependent variable, y, is the lysimeter observations, and the independent variable, x, is the model estimates from the weather data. All calculations, statistics, and graphs were performed in SAS v. 9.1 (SAS Inst. Inc., Cary, NC). In this illustration, the performance tests used are MBE, Pearson correlation coefficient (r), MD plot (Cleveland, 1993), mean squared deviation (MSD), both Lin's (1989) and Liao's (2003) r_c's, e, and D (Section 6.3 in Conover, 1999). The MBE and r are commonly used and so do not need to be described here. The rest are described in the following paragraphs.

Mean-Difference Plot

This plot has been in use at least since Bland and Altman (1986). Most likely, Cleveland (1993) gave it the name *MD-Plot*. The paired differences, $y_i - x_i$, are plotted using the vertical axis scale while the paired means, $0.5(y_i + x_i)$, are plotted using the horizontal axis scale; here i is the observation index with $1 \le i \le n$ where n is the number of paired observations. The MBE is the mean of $y_i - x_i$ over all observations and its standard deviation, standard error, and 95% confidence interval are estimated from the univariate definitions of these statistics applied to the y_i and x_i difference. While conceptually, $-\infty \le \text{MBE} \le \infty$, ideally the MBE ≈ 0. The MD plot is, thus, a way to graphically examine the MBE and its variability. For sound inference, there are several assumptions to assess, particularly the following two: (i) that the paired differences are normally distributed about the MBE and (ii) that the paired differences are not related to the paired means (Bland and Altman, 1986). Generally the zero difference line is included for reference. In this note, as in Fig. 1, an MD plot includes the zero difference line and the MBE with its 95% confidence interval (Meek, 2007). Also in this MD plot, the axes are drawn in a way to visually display the distributions for each variable. Each axis line is drawn as a box-plot to show the extremes, interquartile range, and median. In addition, a histogram is shown to the right of the paired difference axis (vertical) and on top of the paired mean axis (horizontal). Within the area for each histogram the location of the mean is depicted. This plot, thus, helps with the assessment of the underlying assumptions. For example, is the MBE ≈ 0? Is the distribution unimodal? Symmetric? Ideally, a formal univariate analysis that minimally includes the mean, the standard deviation, the skewness, the kurtosis, common quantiles, and a normal distribution test is performed on each variable as well.

Notice, as Lin (1989) points out, the use of MD plot and related analysis alone is not sufficient and can be misleading; hence we suggest using additional analyses and graphs. The suggested additional analyses are two concordance correlation analyses, an associated bivariate (BV) plot (Cleveland, 1993) as modified by Meek (2007), and a formal comparison of the two distributions, the D test. Bivariate plots are not discussed here because they are also commonly used.

Mean Squared Deviation

The MSD can be considered and evaluated as the sum of variance components as follows:

$$\text{MSD} = (\mu_y - \mu_x)^2 + s_y^2 + s_x^2 - 2s_{yx}$$

where μ_y is the mean of the observation variable, μ_x is the mean of the prediction variable with $\mu_y - \mu_x$ another expression for the MBE, s_y is mean deviation y from μ_y, s_x is mean deviation x from μ_x, and s_{yx} is mean cross deviation for y and x (Lin et al., 2002). Notice here that mean deviations and not the standard deviations are used, so the terms are all divided by n, the number of observations, rather than $n - 1$. When the divisor is n, the square root of the MSD can be called root mean square deviation. Sometimes MSD may be estimated with $n - 1$ and referred to as the *lesser bias* form (Lin et al., 2002). Conceptually, the MSD is the sum of squared MBE and its variance. Also notice that $0 \le \text{MSD} \le \infty$. Further discussion on the MSD as well as its relationship to some other common difference measures can be found in Kobayashi and Salam (2000).

Concordance Correlation Coefficient

The MSD components can be included in the related concept of an r_c. To assess the relationship $y = 1x$ by minimizing the squared perpendicular deviation of the paired observations from a 1:1 line (a 45° line, if both the y and x axes scales are identical) under bivariate normal assumption for y and x, Lin (1989) developed an insightful test statistic called the r_c. The r_c statistic is an adjusted version of the well-known Pearson product-moment correlation coefficient, r, and so can be formally evaluated in the same way. The r_c is calculated as

$$r_c = rC_b$$

with C_b (Lin's bias correction factor) = $[(v + 1/v + u^2)/2]^{-1}$. Here $v = s_y/s_x$ and is called a *scale shift* while $u = (\mu_y - \mu_x)/(s_x s_y)^{0.5}$ and is called a *location shift relative to scale*. Notice u is a scaled version of the MBE and v is the equivalent of the ratio of standard deviations because the devisors in s_y and s_x cancel out. A pure location shift could have the data scatter parallel to the 1:1 line through the origin. If $u > 0$, then the scatter is above the 1:1 line. In a pure scale shift, the data scatter would cross the 1:1 line. In general, both shifts are present to some degree. Conceptually, although with some dissent (see Tedeschi, 2006), r_c can be considered the product of precision, r, and accuracy, C_b. Inaccuracy is, thus, due to either one or both u deviating from 0 and v deviating from 1. The bivariate normal and perpendicular error assumptions may be poor in some comparisons. Moreover, ideally, an accuracy measure should include r and the sign of r. Liao (2003) offers an improved r_c that does so by considering the areas of sequential quadrilaterals formed by the 1:1 line and each set of two observations forming adjacent sides of each quadrilateral. The improved accuracy replaces C_b with A_p (Laio's bias correction factor) where, in terms of direct MSD components, A_p is calculated as

$$A_p = \frac{4s_x s_y - r(s_y^2 + s_x^2)}{(2-r)(s_y^2 + s_x^2) + (\mu_y - \mu_x)^2}$$

Alternatively, in terms of location and scale shifts, A_p is

$$A_p = \frac{4\tau - r(1+\tau^2)}{(2-r)(1+\tau^2) + \delta^2}$$

Here, $\tau = s_y/s_x$ or the same as v in the C_b definition, but Liao's location shift relative to scale, $\delta = (\mu_y - \mu_x)/s_x$, is a slightly different form and scaling of the MBE than u is. While conceptually $-1 \le r \le 1$, for modeling purposes, reasonable models will generally have both r and r_c in the interval (0,1) and, ideally, not be different from 1. Interestingly, there are additional analyses related to MSD and r_c that could provide additional insight into model performance, like the total deviation index (Lin et al., 2002).

Nash-Sutcliffe Efficiency

This empirical index is used to indicate agreement between observations (y) and predictions (x). It is included here because it has been widely employed in hydrology and some other related sciences since 1970 (Nash and Sutcliffe, 1970). The definition of e (Nash-Sutcliffe Efficiency) follows:

$$e = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \mu_y)^2}$$

While the range of e is $-\infty \le e \le 1$, reasonable models will have $e > 0$; ideally, $e \approx 1$.

Kolmorgov-Smirnov Statistic

In the BV plot that shows the distributions of y and x, a graphical comparison of them is possible. Let the distributions be identical. If the graph is plotted with the units in equal length scales on both axes so the 1:1 line is 45° then the quantiles and histograms would superimpose on each other exactly if the plot were folded along the 1:1 line. For a reasonable model, hence, assessment of a lack-of-difference between the y and x distributions is primary; there are many ad hoc and formal tests available. Some ad hoc indications of a difference between the x and y distributions include $u \ne 0$, $\delta \ne 0$, and $v \ne 1$. Alternatively, there are formal tests available for this purpose. A significant bias (MBE ≠ 0) is one possibility which reveals a specific dissimilarity. The Kolmorgov-Smirnov two-sample test, D, is another more general test. It is a well-known nonparametric statistic that has been used in both deterministic and stochastic modeling (Tedeschi, 2006). Assume the y and x data are ordered and let S_y and S_x be their respective empirical distribution functions (EDF). The D statistic is then defined as follows:

$$D = \sup_j |S_y(j) - S_x(j)| .$$

Here, sup is mathematical notation for the supremum, and j spans the range of evapotranspiration values. Simply put, the expression means that D is the greatest absolute distance between the two EDF's; hence, its range is $0 \le D \le 1$. The advantage, when using D for a two-sided test, is that it is consistent in revealing all types of differences that may exist between individual y and x distributions. While D can be

Table 1. Evapotranspiration model performance statistics.†

	Concordance Correlation Coefficient Analysis								Efficiency	Difference Measures		Klm.-Smr.
	Lin's Method					Liao's Method						
Model‡	r§	C_b	r_c	u	v	A_p	r_c	δ	e	MBE	MSD	D¶
	Precision	Accuracy		Location	Scale	Accuracy		Location		mm	mm^2	
CIMIS	0.980	0.995	0.975	0.025	1.10	0.991	0.970	0.004	0.954	0.055 ± 0.069	0.243	0.120
										($p \leq 0.428$)#		($p \leq 0.864$)
STAB	0.982	0.963	0.946	0.215	1.19	0.949	0.932	0.033	0.908	0.452 ± 0.077	0.502	0.280
										($p \leq 0.001$)		($p \leq 0.040$)
Change	−0.002	0.032	0.029	−0.190	0.09	0.042	0.038	-0.029	0.046	−0.397	−0.259	

† r, Pearson product-moment correlation coefficient; r_c, concordance correlation coefficient; C_b, Lin's accuracy; u, Lin's location shift relative to the scale shift; v, Lin's scale shift; A_p, Liao's accuracy; δ, Liao's location shift relative to scale; e, Nash-Sutcliffe efficiency; MBE, mean bias error ± standard error of the MBE; MSD, mean square deviation; D, Kolmogorov-Smirnov two-sample test.

‡ There are 50 observations in the data set. The CIMIS model (Model-1) has the wind function based on Pruitt and Doorenbos (1977). The STAB model (Model-2) has the wind function based on Brutsaert (1982).

§ The $p = 0.05$ tabled value for r is 0.23 for the given n.

¶ The $p = 0.05$ tabled value for D is 0.272 for the given n.

The p values are the probability of a Type I error.

determined either analytically or graphically, its p value must be calculated with simple or asymptotic formulae, as is done in the SAS procedure, or checked against tables such as Table A19 in Conover (1999).

RESULTS

Model performance results are summarized in Table 1. The standard errors and Type 1 probabilities for the MBE = 0 are reported in Table 1 to show their connection with the graphical analyses depicted in the MD plots (Fig. 1 and 2). Analogously, the Type 1 probabilities for D along with the concordance analyses results are reported in Table 1 to formally compare them with the distributions depicted in the BV plots (Fig. 3 and 4). Notice that both models are biased low with respect to the measures (u, δ, MBE > 0). Figure 1, the MD plot for Model-1 (CIMIS), shows a reasonably small and acceptable model bias, MBE = 0.055 ($p \leq 0.86$). The related BV plot (Fig. 3) shows reasonable concordance with small location shift ($u \leq |0.1|$, notice the means $\mu_y = 6.48$ and $\mu_x = 6.42$ are similar) and small scale shift ($v - 1 \leq |0.1|$), indicating an accuracy that is likely

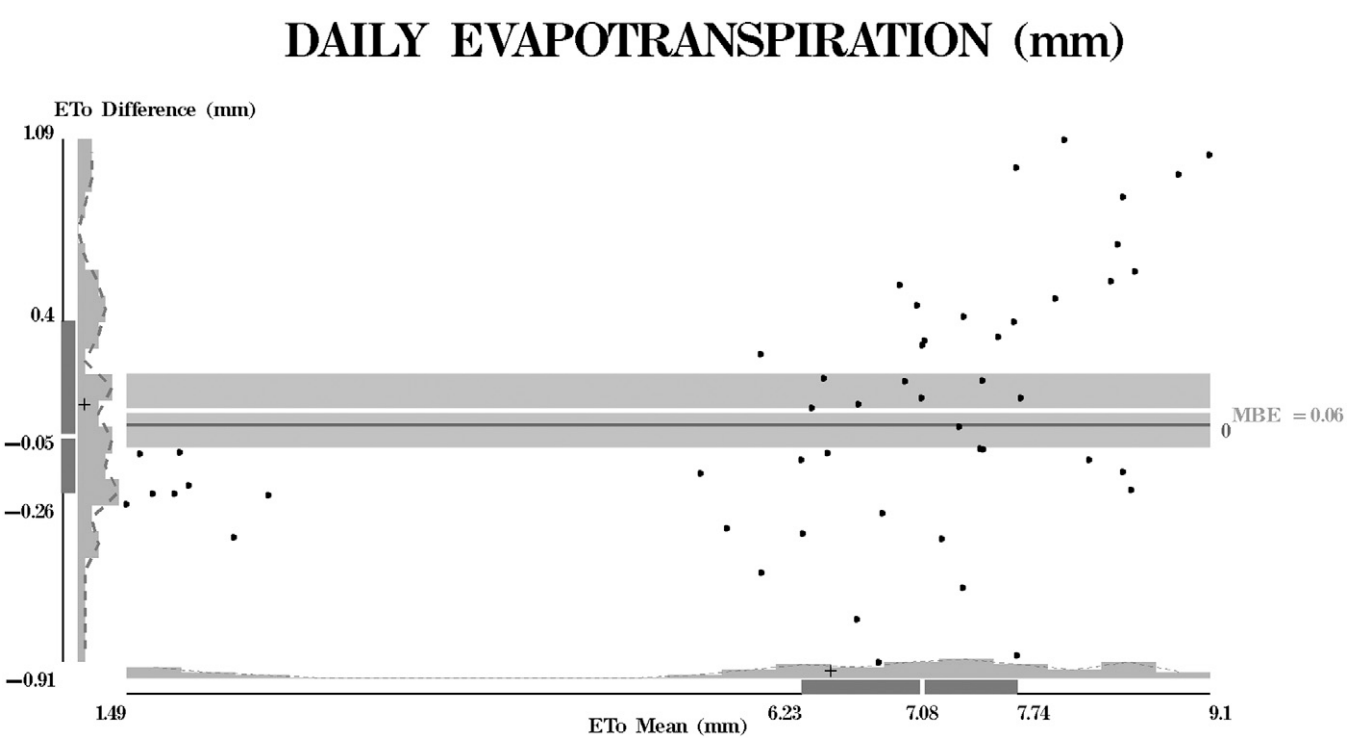

Fig. 1. A mean-difference plot with box-plot axes for the paired means and differences of the Lysimeter measures and the Model-1 (CIMIS) estimates; note that ET_0 denotes the daily evapotranspiration in millimeters. Model-1 is based on the operational model employed in the California Irrigation Management Information System automated weather station network. Both of the thin axis lines span the range of the data. The thick gray bars span the interquartile range with the gap at the median. A simple histogram of each univariate variable is shown for each axis; one just to the right of the vertical axis for the paired differences and one just above the horizontal axis for the paired means. The "+" within each histogram area is the variable's mean value. This mean-difference plot includes black dots to represent the data, the gray band is the 95% confidence interval for the mean bias error (MBE), the gap in the band is the MBE, and the black line is the zero reference line. Notice that the MBE is not different from 0 ($p \leq 0.05$).

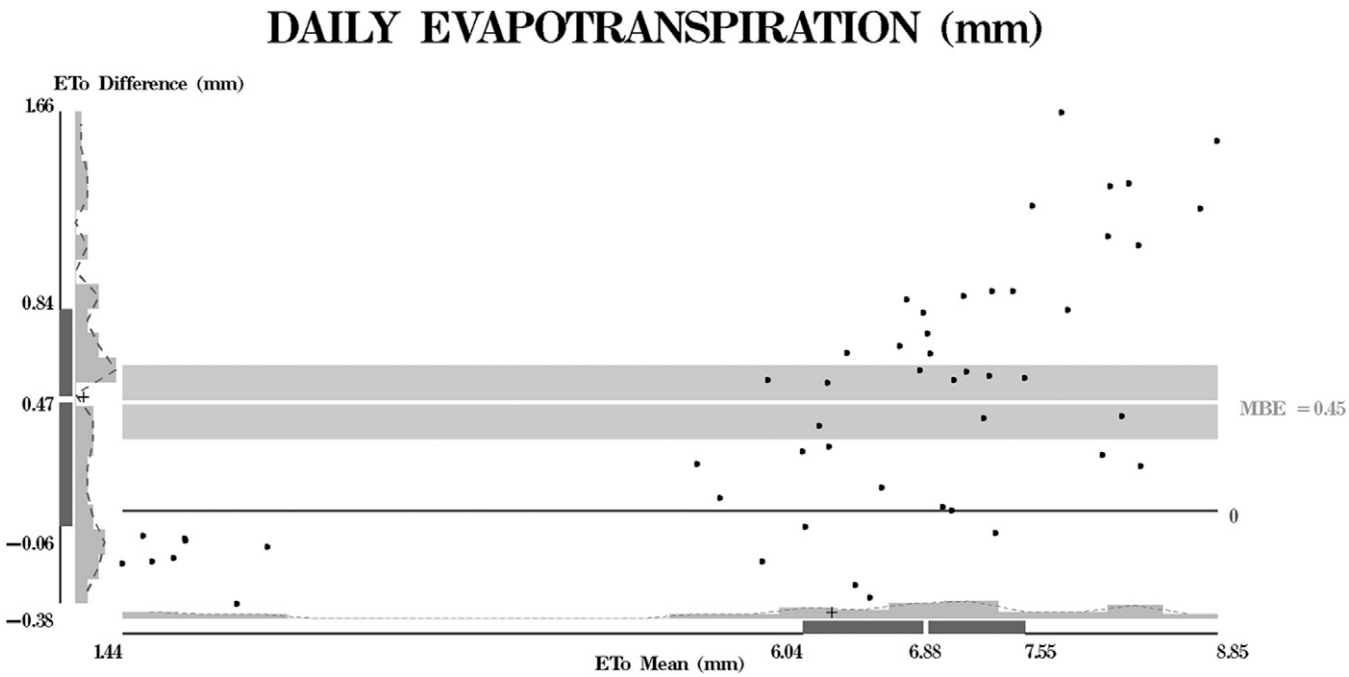

Fig. 2. A mean-difference plot with box-plot axes for the paired means and differences of the Lysimeter measures and the Model-2 (STAB) estimates; note that ET_0 denotes the daily evapotranspiration in millimeters. Model-2 is based on a published aerodynamic stability routine. Reference features are the same as in the Fig. 1 mean-difference plot. Notice that the mean bias error (MBE) is different from 0 ($p \leq 0.05$) and that there is a trend in the paired differences with respect to the paired means.

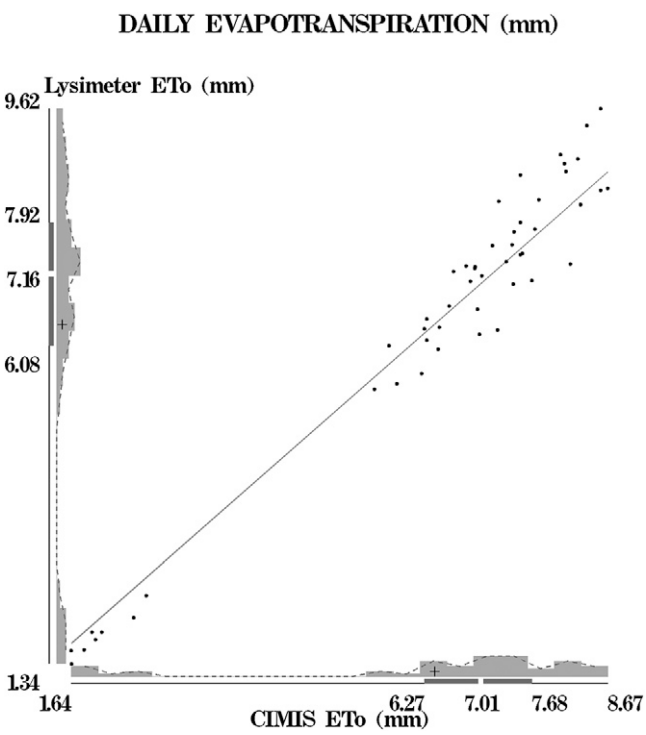

Fig. 3. Model-1 (CIMIS) bivariate plot with box-plot axes similar to those in Fig. 1. Model-1 is based on the operational model employed in the California Irrigation Management Information System automated weather station network; note that ET_0 denotes the daily evapotranspiration in millimeters. Also included is a gray 1:1 reference line. All the related performance measures in Table 1 along with the Fig. 1 mean-difference plot suggest reasonable agreement between the lysimeter observations and the CIMIS (Model-1) model predictions.

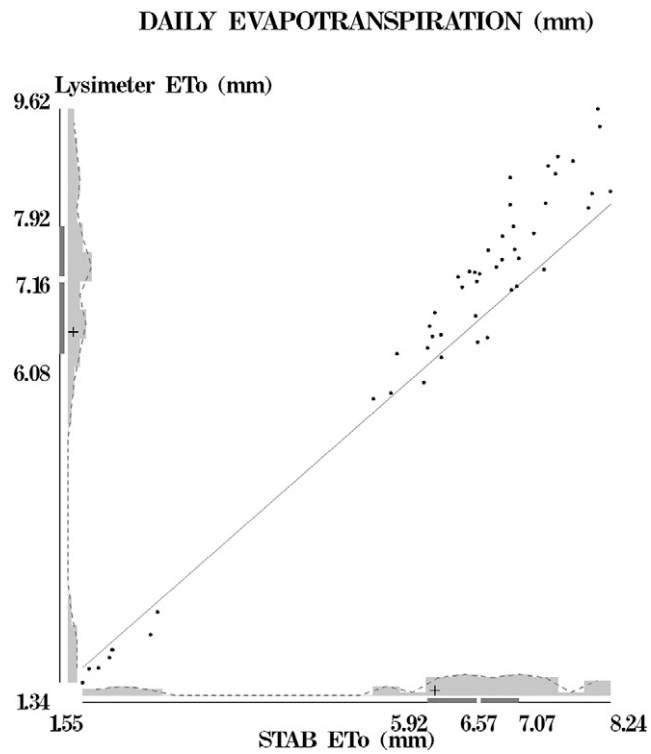

Fig. 4. Model-2 (STAB) bivariate plot with box-plot axes similar to those in Fig. 1. Model-2 is based on a published aerodynamic stability routine; note that ET_0 denotes the daily evapotranspiration in millimeters. This graph, along with the Fig. 2 mean-difference plot and the related performance measures in Table 1, indicate a considerable lack of agreement between the lysimeter observations and the STAB (Model-2) predictions.

generally acceptable to most users. Multiple distribution tests for the difference distribution indicate reasonable normality; these results are also omitted for brevity. In contrast, while the precision (r) for Model-2 (STAB) is slightly better than that for Model-1 (0.982 vs. 0.980), both of the related accuracy components are considerably worse. Figure 2 shows extreme model bias (MBE = 0.452 [$p \leq 0.04$], over 8 times that of Model-1) and a pronounced systematic relationship with the paired means. The related BV plot (Fig. 4) shows lower concordance with the comparatively poorer accuracy due both to a 10-fold larger location shift and about 10% larger scale shift, $u = 0.215$ (notice the means $\mu_y = 6.48$ and $\mu_x = 6.03$ are not similar) and $v - 1 = 0.19$. In addition, multiple distribution tests for the difference distribution only marginally suggest normality (again, the details are omitted for brevity). The differences in both Liao's A_p (0.991 vs. 0.949) and r_c (0.970 vs. 0.932) are larger still than those for Lin's C_b (0.995 vs. 0.963) and r_c (0.975 vs. 0.946) analogs. Moreover, the other performance measures also indicate that Model-2 estimates are probably unacceptable to most users. For example, the MSD (smaller is better) for Model-2 is over twice that of Model-1 (0.502 vs. 0.243), while e (closer to 1 is better) for Model-2 is about 5% smaller than e for Model-1 (0.908 vs. 0.954). Finally, the D test result, $D = 0.12$, suggests similarity between the distribution for the observations and that for the CIMIS (Model-1) estimates because $p \leq 0.86$ is not significant at a reasonable level. For the STAB model (Model-2) comparison, however, the situation reverses because $p \leq 0.04$ for $D = 0.28$; this maximum EDF deviation occurs at about 7 mm of evapotranspiration.

DISCUSSION AND CONCLUSIONS

Using a physically based model over an empirical model is philosophically desirable. Practically speaking, however, the former should also perform better in terms of any chosen set of measures, target values for the measures, or concepts to be selected. Here, the STAB model does not perform better for any measure but r. Perhaps with further work it could be improved, but such a goal is beyond the purpose and scope of this work. Sound model development and assessment includes more than setting goals for performance measures or comparing them. It should include a careful conceptual analysis tailored to the purpose of the application (Oreskes et al., 1994; Tedeschi, 2006). In addition, it is unreliable to base a decision of model acceptance on a single performance measure, let alone on the evaluation of a single data set. None-the-less, given the appropriate careful considerations, this example points out the merits of the MD plot along with concordance correlation analysis and distribution comparisons. Any of the difference measures or plots, especially the MD plot, reveals the serious bias in Model-2 estimates; however, they do not directly reveal or deal with the scale shift. An r value alone does not adjust for either location or scale shift, but either one of the r_c statistics does. In addition, the comparative concordance evaluation can be cast in terms of both comparative precision, accuracy, and the related distribution and probability differences. The modeler's toolbox should therefore include MD plots, both concordance correlation analyses, and distribution procedures for model performance assessment because, together, they provide simple and sound statistically based tests that can also add analytical insight.

As previously mentioned, there are other related performance measures that also merit consideration, like the total deviation index. In addition, Legates and McCabe's (1999) robust versions of d and e are simple and straightforward calculations and may be of interest to many modelers; hence, they should also be in a modeler's toolbox. Robust versions of the other measures are possible and perhaps should be further considered; such measures require a great deal of caution and care in use for the model performance application. Poor predictions due to model bias or inadequacy are the reason for assessing a model's performance and, hence, should not be subjected to robust procedures. Mistakes and other problems with observations, however, should be identified and screened out or down weighted.

ACKNOWLEDGMENTS

This work was supported by the USDA-ARS National Soil Tilth Laboratory, Ames, IA, Dr. J.L. Hatfield, Director. Thanks to J.W. Singer, USDA-ARS, Ames, IA; R.P. Ewing, ISU, Ames, IA; and the journal reviewers for their valuable comments. This manuscript is dedicated to the memory of Wm. O. Pruitt, a long-term researcher and lecturer at the University of California at Davis. Bill Pruitt demonstrated the highest ethical and integrity research standards. Additionally, he provided a legacy for future generations of ET students at UC Davis.

REFERENCES

Bland, J., and D. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical agreement. Lancet I:307–310.

Brutsaert, W. 1982. Evaporation into the atmosphere: Theory, history, and applications. D. Reidel Publ., Boston, MA.

Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, NJ.

Conover, W.J. 1999. Practical nonparametric statistics. J. Wiley & Sons, New York.

Fila, G., G. Bellocchi, M. Donatelli, and M. Acutis. 2003. IRENE_DLL: A class library for evaluating numerical estimates. Agron. J. 95:1330–1333.

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.

Howell, T.A., R.L. McCormick, and C.J. Phene. 1985. Design and installation of large weighing lysimeters. Trans. ASAE 28:106–112, 117.

Howell, T.A., D.W. Meek, C.J. Phene, K.R. Davis, and R.L. McCormick. 1984. Automated weather data collection for research on irrigation scheduling. Trans. ASAE 27:386–391, 396.

Joyce, A., J. Adamson, B. Huntley, T. Parr, and R. Baxter. 2001. Standardisation of temperature observed by automatic weather stations. Environ. Monit. Assess. 68:127–136.

Kobayashi, K., and M.U. Salam. 2000. Comparing simulated and measured values using mean squared deviation and its components. Agron. J. 92:345–352.

Legates, D.R., and G.J. McCabe, Jr. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35:233–241.

Liao, J.J.Z. 2003. An improved concordance correlation coefficient. Pharm. Stat. 2:253–261.

Lin, L.I. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268.

Lin, L.I., A.S. Hedayat, B. Sinha, and M. Yang. 2002. Statistical methods in assessing agreement: Models, issues, and tools. J. Am. Stat. Assoc. 97:257–270.

McGinn, S.M., K.A. Beauchemin, A.D. Iwaasa, and T.A. McAllister. 2006. Assessment of the sulfur hexafluoride (sf6) tracer technique for measuring enteric methane emissions from cattle. J. Environ. Qual. 35:1686–1691.

Meek, D. 2007. Two macros for producing graphs to assess agreement between two variables, paper D6. [CD-ROM] *In* A. Katschke (ed.) Data Visualization and Graphics Section, 2007 MidWest SAS Users Group Annual Conference Proc., Des Moines, IA. 28–30 Oct. 2007. SAS Inst., Cary, NC.

Nash, J.E., and J.V. Sutcliffe. 1970. River flow forecasting through conceptual models, part I. J. Hydrol. 10:282–290.

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

Phene, C.J., D.W. Meek, R.L. McCormick, and K.R. Davis. 1986. Calibration and use of Penman's equation for irrigation scheduling. ASAE Paper No. 86–2594. ASAE, St. Joseph, MI.

Pruitt, W.O., and J. Doorenbos. 1977. Empirical calibration, a requisite for evapotranspiration formulae based on daily or longer mean climatic data? Intl. Round Table Conf. on Evapotranspiration, Budapest, Hungary. Int. Commission on Irrigation and Drainage, New Delhi, India.

Snyder, R., W.O. Pruitt, D.W. Henderson, and A. Dong. 1985. California irrigation management information system. LAWR Paper 10013-A. Final Rep. Univ. of California, Davis.

Tedeschi, L.O. 2006. The assessment of the adequacy of mathematical models. Agric. Syst. 89:225–247.

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

Willmott, C.J. 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63:1309–1313.

Willmott, C.J. 1984. On the evaluation of model performance in physical geography. p. 184–194. *In* G.L. Gaile and C.J. Willmott (ed.) Spatial statics and models. D. Reidel Pub. Co., Boston, MA.

Willmott, C., S. Ackleson, R. Davis, J.J. Feddema, K. Klink, D. Legates, J. O'Donnell, and C. Rowe. 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90:8995–9005.