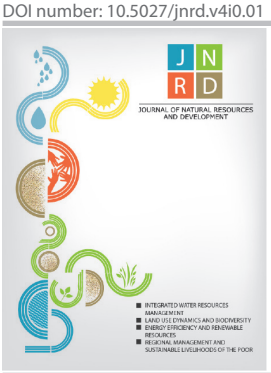
DOI number: 10.5027/jnrd.v4i0.01



JOURNAL OF NATURAL RESOURCES AND DEVELOPMENT

A new novel index for evaluating model performance.

M. H. Ali*, I. Abustan

School of Civil Engineering, University Sains Malaysia

* Corresponding author : mha_bina@yahoo.com, hossain.ali.bina@gmail.com. Permanent address: Agricultural Engineering Division, Bangladesh Institute of Nuclear Agriculture

Article history

Received 27.01.2013
Accepted 11.06.2013
Published 27.01.2014

Keywords

Model evaluation
Statistical indicators
Model performance

Abstract

A vast array of scientific literature is concerned with simulation models. The aim of models is to predict the unknown situation as close to as real one. To do this, models are validated and examined for their performance under known condition. In this paper, commonly used model performance evaluation indices are overviewed and examined under different situations. Difference based, efficiency based (Nash and Sutcliffe coefficient, model efficiency of Loague and Green, Legates and McCabe's index) and composite indices (such as index of agreement, *d*, and d_r) were found ambiguous, inconsistent and not logical in many cases. A new index, Percent Mean Relative Absolute Error (PMRAE), is proposed which is found unambiguous, logical, straight-forward, and interpretable; thus can be used to evaluate model performance. The model evaluation performance ratings based on PMRAE are also suggested.

Introduction

A range of simulation models and decision support systems have been developed and are being used for several decades in different fields. Simulation models have been successfully used to provide simulations of crop growth and development (Geerts et al. 2009, Stockle et al. 2003), hydrologic variables (Suleiman 2008), water and solute transport (Crescimanno and Garofalo 2005,Dust et al. 2000), solar radiation (Rivington et al. 2005), environmental impacts (Stockle et al. 1992) and many other areas. One important aspect in the model development process is the model evaluation. Model outputs are compared /examined with observed (or known) data gathered under respective conditions, both by quantitative and graphical methods.

Various statistical and efficiency-based indices/indicators and test statistics have been suggested and used by diffident model developers and users to judge the model performance. Among these, recommendation by Nash and Sutcliffe (1970), Fox (1981), Willmott (1982, 1985), and Loague and Green (1991) are prominent.

Among statistical indices, some of them quantify the departure of the model output from observed or experimental measurements, while others focus on correlation between model predictions and measurements. In essence, Fox (1981) recommended that the following four types of difference measures should be calculated and reported: mean error, mean absolute error, variance of the distribution of difference, and root mean square error (or its square - the mean square error).These difference-based statistics quantify the departure of the model outputs from the measurements.

Indicators for specific fields are also suggested. Bellocchi et al. (2002) proposed a fuzzy expert system to calculate a composite indicator for performance evaluation of solar radiation. They used

correlation coefficient (r), relative root mean square error (RRMS), model efficiency (EF), and *t*-Student probability to make aggregated form. Confalonieri et al. (2010) proposed a fuzzy-based, indicator for evaluation of soil water content simulation. Jacovides and Kontoyiannis (1995) proposed mean bias error (MBE) and root mean square error (RMSE) in combination with the t-statistic as statistical indicators for the evaluation and comparison of evapotranspiration computing models.

Among the difference and/or statistical measures, mean error (ME), root mean square error (RMSE), relative error (RE), and correlation coefficient (r) are widely used in different fields –crop growth and yield (Geerts et al. 2009), irrigation scheduling (Liu et al. 1998), hydrological (Shen et al. 2009), environmental (Wagener and Kollat 2007), solar radiation (Rivington et al. 2005), pollution simulation model (Yang et al. 2007), etc. Model efficiency (EF) is used in almost every field of simulation. The above indices are used for both single model evaluation and comparison of multiple models (Prasher et al. 1996). Martorana and Bellocchi (1999) identified the mean squared error of prediction as the fundamental statistical index on which other widely used squared differences are based. While Willmott and Matsuura (2006) noted that RMSE is an inappropriate measure of average error because it is a function of three characteristics of a set of errors, rather than of one (average error).

Yang et al. (2000) evaluated different statistical methods to evaluate crop-nitrogen simulation model, N_ABLE. They suggested that two sets statistics can be used: (a) mean of error (ME), root mean square error (RMSE), forecasting efficiency, and paired t-statistic; (b) ME, mean absolute error, forecasting coefficient, and F-ratio of lack of fit over experimental error. They noted that either set can give the same conclusions which could not be quantitatively detected by graphical method. The use of test statistics (e.g. F, t-test, etc.) to judge the error variance between observed and simulated outputs have the possibility of producing type-I or type-II error.

Willmott (1981) demonstrated that the correlation coefficient, r (Pearson's product-moment correlation coefficient) can be misleading measure of accuracy – 'r' between very dissimilar model-predicted variable and observed one can easily approach 1.0. Willmott (1982) discussed other drawbacks of 'r' and 'R^2', and proposed an "index of agreement (d)". He noted that the index 'd' is intended to be a descriptive measure, and it is both a relative and bounded measure which can be widely applied for cross-comparisons between models. Willmott et al. (2011) suggested a refined index (d_r) considering the problem of *d*.

Among the efficiency-based indices (EF) suggested for model performance evaluation, widely used ones are Nash and Sutcliffe coefficient (Nash and Sutcliffe 1970) and model efficiency of Loague and Green (1991). Many researchers (Addiscott and Whitmore 1987, Martorana and Bellocchi 1999, Rivington et al. 2005, Moriasi et al. 2007) noted that a model may be judged suitable according to one statistic but it may be deficient according to another statistic. Alexandrov et al. (2011) emphasized the need of standardized evaluation tool.

The purpose of this paper is to examine all of the above indices, and suggest a logical, stable, unambiguous and straight-forward index for model performance evaluation.

Materials and methods

Definition of commonly used statistical measures and indices for model performance evaluation

Before going to analyze the indices, it would be useful to define them along with their perspectives. So, they are described below. For synchronization of all the indices, observed or measured value is designated by O_i, and predicted or simulated value is designated by P_i, although the original proposed symbol may be different in some cases.

Difference based Statistical indicators

(i) Mean bias or Mean error (ME)(Fox 1981):

$$ME = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i) \quad (1)$$

Where, N is the number of observations.

(ii) Mean Absolute error (MAE) (Fox 1981):

$$MAE = \frac{1}{N}\sum_{i=1}^{i=N}|P_i - O_i| \quad (2)$$

(iii) Root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \quad (3)$$

The RMSE quantifies the dispersion between simulated and measured data. Ideally, the value of ME, MAE, and RMSE should be zero.

iv) Relative error (RE) or relative root mean square error (RRMSE) (Loague and Green 1991, Bellocchi et al. 2002):

$$RE = \frac{RMSE}{\overline{O}} \times 100 \quad (4)$$

Where, $\overline{O}$ is the mean of observed values. The RE may vary from 0 to positive infinity. The smaller the RE is, the better the model performance. Sometimes it is expressed as percentage form.

v) Scaled Root-mean-Square-Error (SRMSE) (Dust et al. 2000):

$$SRMSE = \frac{1}{\overline{O}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \quad (5)$$

In essence, the RE and SRMSE are the same.

Efficiency based indicators

(i) Nash and Sutcliffe Coefficient of efficiency (E_{NS}) (Nash and Sutcliffe 1970):

$$E_{NS} = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \quad (6)$$

Nash-Sutcliffe coefficient of efficiency (E_{NS}) varies between - ∞ and 1.0, and E_{NS}=1 is the optimum value. The E_{NS}≤0.0 indicates unsatisfactory performance, and $0<E_{NS}<1$ is considered as the acceptable range.

(ii) Model efficiency of Loague and Green (E_{LG}) (Loague and Green 1991):

$$E_{LG} = \frac{\sum_{i=1}^{N}(O_i - \overline{O})^2 - \sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \quad (7)$$

An ideal value of E_{LG} is unity. Its upper limit is 1, and lower value can negative infinity.

The Nash-Sutcliffe coefficient of efficiency (E_{NS}) and the model efficiency of Loague and Green (E_{LG}) are the same. So, only one will be discussed in the later section.

(iii) Legates and McCabe's index (E_{LM}) (Legates and McCabe 1999) Legates and McCabe's index (E_{LM}) is written as:

$$E_{LM} = 1 - \frac{\sum_{i=1}^{N}Abs(P_i - O_i)}{\sum_{i=1}^{N}Abs(O_i - \overline{O})} \quad (8)$$

Other composite indicators

(i) Index of Agreement (d) (Willmott 1982):

$$d = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}[O'_i + P'_i]^2}, \quad 0 \le d \le 1 \quad (9)$$

where $O'_i = |O_i - \overline{P}|$, $P'_i = |P_i - \overline{P}|$, O_i is the observed value, P_i is the simulated value and $\overline{P}$ is the simulated mean.

(ii) Refined index of Willmott et al. (2011)

The refined index of Willmott et al. (2011) (d_r) can be written as:

$$d_r = 1 - \frac{m_1}{c*m_2}, \quad \text{when } m_1 \le c*m_2 \quad (10)$$

$$= \frac{c*m_2}{m_1} - 1 \quad \text{when } m_1 > c*m_2$$

Where,

$$m_1 = \sum_{i=1}^{N}Abs(P_i - O_i)$$

$$m_2 = \sum_{i=1}^{N}Abs(O_i - \overline{O}_i) \quad \text{and } c = 2$$

Proposed new index

Percent mean absolute relative error (PMARE)

It is the 'mean absolute relative error', expressed in percentage.

$$PMARE(\%) = \frac{100}{n}\sum_{i=1}^{n}\frac{Abs(O_i - P_i)}{O_i} \quad (11)$$

Where, '*Abs*' indicates absolute value (of the difference between observed and simulated value). Theoretically, the value of PMARE ranges from 0% to ∞ (positive infinity). The interpretation and characterization of the index are discussed later.

Data for comparison of indices

To test the statistics and indices, both the field observed data and simulated random data were used.

Simulation comparison with field observed data

Field data are originated from wheat experiment, where diverse irrigation treatments were applied representing different strategies of deficit irrigation. Simulation was performed using AquaCrop model of FAO (Steduto et al. 2009). Before simulation, calibration of the model was performed using one year data. The model AquaCrop produces inferior simulations at extreme dry condition (herein referred as 'odd simulation'– sometimes referred in the literature as 'outliers'), which is a common problem in many models. Observed and simulated outputs are summarized in Table 1, which are used to explore the behavior of the indices.

Table 1. Observed and simulated yield of wheat grain & total biomass

Data year	Treatment/ Sl.no.	Grain yield (t/ha)		Total biomass yield (t/ ha)	
		Observed	Simulated	Observed	Simulated
1st	1	2.071	1.293	7.06	6.614
	2	3.978	3.956	11.649	11.384
	3	3.721	3.956	10.351	11.383
	4	3.872	3.779	10.643	10.962
	5	3.859	3.734	11.197	10.887
	6	3.846	3.586	10.946	10.649
	7	3.739	3.191	10.276	9.741
	8	3.618	3.734	10.227	10.886
	9	4.017	4.015	11.85	11.473
	10	3.281	1.707	9.588	7.384

Table 1. Observed and simulated yield of wheat grain & total biomass (continuation)

Data year	Treatment/ Sl.no.	Grain yield (t/ha)		Total biomass yield (t/ ha)	
		Observed	Simulated	Observed	Simulated
2nd	1	1.574	0	4.246	0
	2	3.404	3.802	10.50	11.151
	3	3.144	3.798	11.223	11.142
	4	3.169	3.688	10.366	10.854
	5	3.168	3.613	10.145	10.694
	6	3.395	3.271	10.265	10.05
	7	3.141	2.901	8.991	9.281
	8	2.994	2.567	9.24	9.004
	9	3.48	3.802	11.61	11.151
	10	2.779	1.519	9.045	7.167

Simulation comparison with Random data

To show the behavior of the indices under different patterns of data series, values of *O* and *P* were created (generated) using a random data generator. More specifically, 3 sets of random data of size n=20 were generated separately for *O* and *P* using a random number generator (RANDOM.ORG, 2012) [Table 2, Fig.1]. The randomness comes from atmospheric noise.

Calculation of the indices

The indices were calculated using Microsoft spreadsheet following the equations mentioned earlier.

Table 2. Data sets (Random numbers) generated using 'Radom Number Generator'

Sl no.	Set-1		Set-2		Set-3	
	Observed	Simulated	Observed	Simulated	Observed	Simulated
1	33	46	11	24	17	14
2	50	23	12	11	41	19
3	36	35	44	43	7	35
4	33	45	36	8	41	4
5	25	21	20	20	27	18
6	7	22	17	16	36	24
7	2	23	33	50	29	26
8	42	43	30	8	36	7
9	17	18	38	26	21	5
10	18	42	17	42	10	32
11	45	12	21	16	42	38
12	23	39	8	12	39	27
13	21	7	40	46	28	32
14	42	44	20	15	30	6
15	47	13	18	12	43	47
16	7	7	32	24	46	41
17	11	34	16	1	10	33
18	20	15	18	5	42	13
19	29	23	32	31	5	46
20	12	22	21	10	46	17

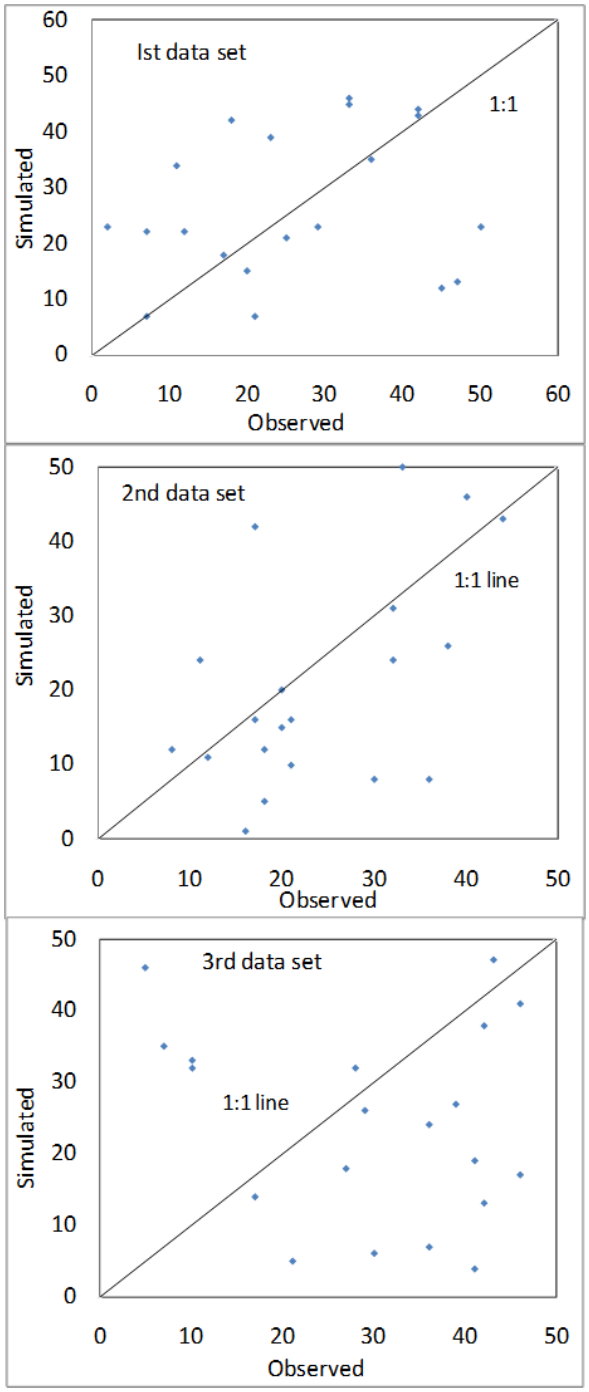


Figure 1. Pattern of observed versus simulated random data sets

Results and Discussions

Grain yield of wheat

The statistical parameters and indices under different conditions ("with" and "without" odd simulated values) are presented in Table 3. The data points (with odd values) are graphically illustrated in Fig.2 along with 1:1 line.

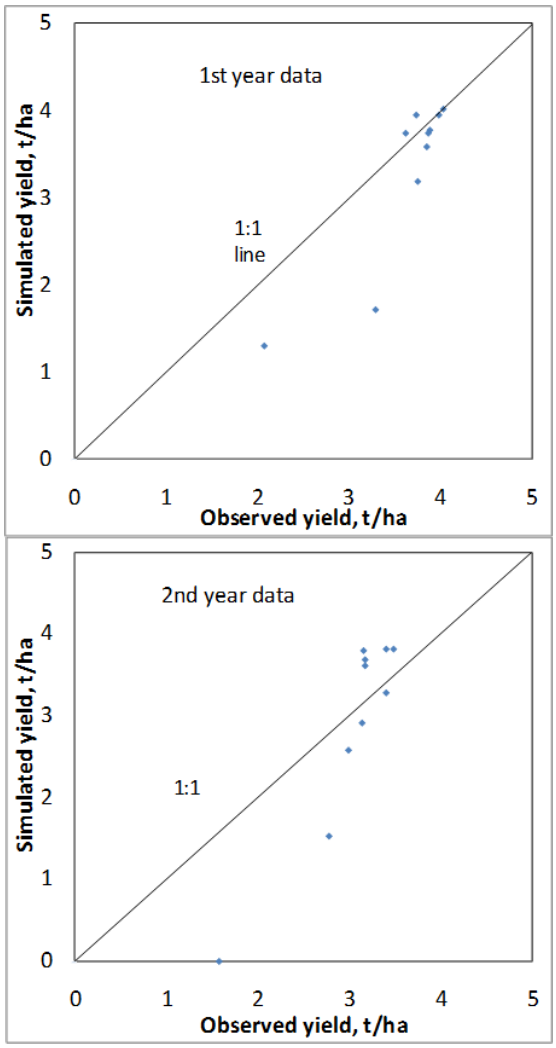


Figure 2. Pattern of observed versus simulated grain yield of wheat

For the simulation year-1, while the 'odd simulated' values are omitted from the calculation, the difference-based statistical indicators – mean error (ME), mean absolute error (MAE), root mean square error (RMSE), and relative error (RE) decreased compared to those with 'odd simulated values'; which is logical. The efficiency based indicators – E_{NS} or E_{LG}, E_{LM}, index of agreement *d*, and new index of agreement d_r decreased; but they should be increased. Similar behaviors are also observed for the year-2.

For the combined data, the statistical indicators followed the logical behavior. Here, the E_{NS} and E_{LM} followed the logical trend – higher values for 'without odd data' (i.e. with good simulated data). But the *d* followed the reverse behavior – decreased with good simulated values. The PMARE always followed the logical behavior, and no ambiguous result.

From the different data sets, it is revealed that the difference-based statistical indicators gave consistent and logical measures. The behavior of E_{NS} and E_{LM} is inconsistent, and reverse in two cases. The behavior of *d* is reverse in all the studied cases. Similar behavior is also noticed by 'r'.

The behavior of E_{NS} and E_{LM} may be due to their inherent formulation/structure. From the equation of E_{NS} and E_{LM}, it is revealed that they are more dependent on observation range (O_i and O) than the difference between the observed and predicted values. Thus, the E_{NS} and E_{LM} are more sensitive to observed range/fluctuation. Hence the output is not consistent and reliable. Similar behavior is also noticed by *d*. In the studies cases, the outputs are consistently reverse to the logical direction. For d_r, the behavior is inconsistent for 2 cases – 1st& 2nd year data.

Case of total biomass yield

The observed and simulated total biomasses are illustrated in Fig. 3 along with 1:1 line. The statistical parameters and efficiency indices are presented in Table 4. The *r* value shows reverse behavior (opposite

Table 3. Statistical and efficiency indicators for evaluating simulation performance of wheat grain yield under different conditions

Statistical indicator	1st year		2nd year		Combined data	
	With odd simulations	Without odd simulations (excluding no. 1 & 10)	With odd simulation	Without odd simulation (excluding no. 1 & 10)	With odd simulations	Without odd simulations
Mean Bias (t/ha)	-0.305	-0.087	-0.129	0.193	-0.217	0.053
Mean absolute bias (MAE) (t/ha)	0.375	0.175	0.596	0.391	0.486	0.283
RMSE (t/ha)	0.595	0.240	0.740	0.420	0.672	0.342
RE (%)	16.54	6.268	24.47	12.98	20.28	9.68
Pearson's moment correlation coefficient (r)	0.887	0.440	0.930	0.581	0.867	0.572
E_{NS} or E_{LG} (%)	-18.44	-269.59	-101.14	-612.66	-22.40	-7.93
E_{LM}	-0.0152	-0.685	-0.726	-1.753	-0.081	0.0473
Index of agreement (d)	0.951	0.534	0.809	0.611	0.838	0.774
New index of agreement, d_r	0.492	0.158	0.14	-0.273	0.459	0.524
PMARE (%)	12.27	4.65	24.3	12.22	18.3	8.43

to logical, MAE & PMARE) for 2nd year & combined data - higher value for "with odd simulation'.

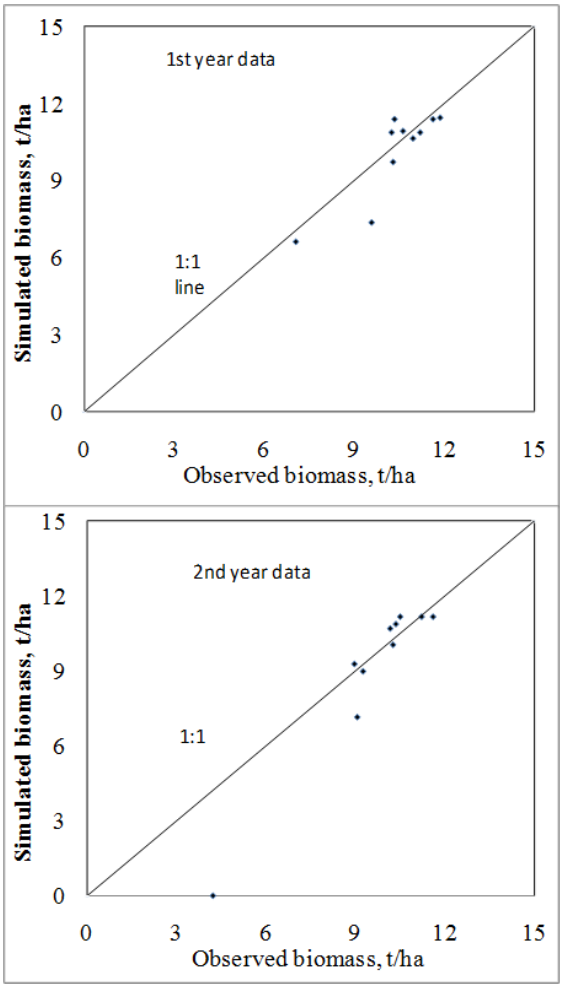


Figure 3. Pattern of observed versus simulated biomass in different years

Behavior with Random data

The values of statistical and efficiency based indices for 3 random data sets, with 'all data' (herein referred as 'with odd simulation') and 'without extreme values' (2 extremes) (herein referred as 'without odd simulation') are summarized in Table 5. Here, the E_{NS} and d_r do not follow the logical behavior of difference-based measures MAE and RMSE (and also PMARE) for the 1st & 3rd data sets. The E_{LM} shows reverse trend for 3rd data set. The results indicate that the *r*, E_{NS}, E_{LM}, and d_r show ambiguous performance rating under different conditions.

Discussion

Now-a-days, the modeling and the use of models are becoming the major thrust in all branches of science. It is increasingly important that discussion of model evaluation procedure to be expanded in order that logical, consistent and generally accepted indicator(s) is identified. The indicators should appropriately quantify objective of model evaluation – that is, should direct towards the answer of model usability. It is logical demand that an '*ideal indicator*' for model performance evaluation should:

(i) Have straight-forward physical meaning and interpretation
(ii) Indicate the strength (accuracy) or pit-fall (weakness) of the prediction capability, so that decision can be made regarding usefulness of the model
(iii) Have consistent value/trend with the logical direction, and no ambiguous performance rating

Graphical method gives the overall and real picture, while the different indices give quantitative measures. The diagnosis that can be made from the graph, must be supported by the quantitative measures. The indices should also be consistent in their results. Otherwise, the

Table 4. Statistical and efficiency based indicators for evaluating simulation performance of total biomass yield under different conditions

Statistical indicator	1st year data		2nd year data		Combined data	
	With odd simulations	Without odd simulations (excluding no. 1 & 10)	With odd simulation	Without odd simulation (excluding no. 1 & 10)	With odd simulations	Without odd simulations
Mean Bias (t/ha)	-0.242	-0.024	-0.514	0.123	-0.378	0.045
Mean absolute bias, (MAE) (t/ha)	0.644	0.471	0.909	0.371	0.777	0.424
RMSE (t/ha)	0.857	0.525	1.514	0.413	1.230	0.476
RE (%)	8.26	5.02	15.83	4.02	12.34	4.58
Pearson's moment correlation coefficient (r)	0.87	0.93	0.972	0.885	0.943	0.917
ENS or ELG (%)	55.55	84.30	40.11	75.04	47.92	82.03
E_{LM}	0.266	0.464	0.324	0.413	0.333	0.449
Index of agreement, d	0.910	0.963	0.915	0.934	0.918	0.956
New index of agreement, d_r	0.633	0.732	0.662	0.707	0.667	0.724
PMARE (%)	6.49	4.65	14.96	3.61	10.72	4.16

Table 5. Statistical and efficiency based indicators for evaluating simulation performance based on random data

Statistical indicator	1st data set		2nd data set		3rd data set	
	All data (With odd simulations)	Without odd simulation (excluding 2 extremes)	All data	Without odd simulation (excluding 2 extremes)	All data	Without odd simulations
Mean Bias (ME)	0.70	-1.67	-3.20	-5.66	-5.60	-10.05
Mean absolute bias	0.644	0.471	0.909	0.371	0.777	0.424
(MAE)	13.1	12.11	9.70	8.66	17.80	15.94
RMSE	16.87	16.20	12.68	11.60	21.38	19.26
RE (%)	64.90	57.52	52.40	45.78	71.76	59.38
Pearson's moment correlation coefficient (r)	0.408	0.45	0.271	0.50	-0.546	-0.358
ENS or ELG (%)	-41.22	-50.14	-53.64	-31.34	-155.39	-187.91
ELM	-0.07	-0.05	-0.06	0.05	-0.56	-0.638
New index of agreement, dr	0.467	0.446	0.469	0.526	0.221	0.181
PMARE (%)	108.96	51.12	45.48	35.79	117.61	62.89

particular quantitative index is not suitable for model comparison, and should be abundant from model performance measure.

Legates and McCabe (1999) suggested that correlation and correlation-based measures (e.g. the coefficient of determination, R^2) should not be used to assess the goodness-of-fit of hydrologic or hydro-climatic model, as these measures were found over-sensitive to extreme values (outliers) and are insensitive to additive and proportional differences between model predictions and observation. Willmott (1981) found ambiguous behavior of correlation coefficient,'r'. The present study also showed ambiguous behavior of 'r'.

Within the domain of efficiency based indicators, McCuen et al. (2006) showed that the outliers can significantly influence sample values of the Nash–Sutcliffe efficiency index (E_{NS}). In the present study, E_{NS} and E_{LG} also showed ambiguous result due to the presence or absence of externalities (extreme values).

Willmott et al. (2011) proposed a new index, dr, and they compared the d_r with 'mean absolute error (MAE)' of the data sets, which varies logically with MAE. But this should be compared with mean absolute relative error, because MAE can vary with different data pattern/set, while the 'mean absolute relative error' value may be the same (i.e. no change in relative pattern). In the present study, the dr index does not follow the logical trend within a particular data set, as in Table 2 (combined analysis); and also ambiguous among different sets (1st year and combined data) – with PMARE value. Similar inconsistencies are also observed for random data sets (Table 4, 1st & 3rd data sets – with PMARE).

As the behavior of EF, *d*, d_r and r are not consistent and logical for all cases (ambiguous, conflicting performance rating); they should be avoided from model performance measure.

The 'mean absolute error' (MAE)and 'mean bias error' (ME) have been suggested by Willmott and Matsuura (2006). But the MAE or ME does not tell about the level or degree of error, and the MBE can 'neutralize the amount of error' if the error occurs on both positive and negative directions. The 'mean absolute relative error', when expressed as percentage, that is 'percent mean absolute relative error' (PMARE) (eqn.11), overcomes the above deficiencies. It has merit over 'mean absolute error' that it directly indicates the strength or weakness of the simulation; and thus helps to decide accept or reject the model. Theoretically, the value of PMARE can range from 0% to ∞ (positive infinity). As it is a measure of error (but relative – with respect to observed, which is logical than any other measure), the optimum value is 0.0, indicating no error (that is perfect simulation). Low magnitudes indicate less error (i.e. better model simulation) and the higher values indicate higher error (i.e. less perfect simulation). The 0<PMARE<100 can be considered as the practical/acceptable range. Performance rating based on any indicator may depend on the model type, field of application (i.e. sensitivity of the work/project where the model output will be used), availability of real-world data, etc. In general, for the PMARE value, the following ratings may be used as a guide (Table 6):

Table 6. Suggested performance rating for model evaluation based on new index, PMARE

PMARE value (%)	Model rating
0 - 5	Excellent
5-10	Very good
10 - 15	Good
15 - 20	Fair
20 - 25	Moderate
>25	Unsatisfactory

The above threshold/maximum limit for rating a model is determined/ suggested after examining various data sets, by sequentially omitting the data having higher difference (in percent) between observed and simulated values, and determining PMARE.

Based on the required precision, the user can choose lower PMARE value. On the other hand, where no other means/data are available, the user can use a model having even a higher PMARE value (say, 25%) to get a forecast.

The PMARE has distinct advantages over the other indicators:

(i) It is simple to calculate
(ii) Has direct physical meaning
(iii) Indicates directly the accuracy or pit-fall of the simulation, and thus helps to decide about the acceptability (or usefulness) of the model
(iv) No ambiguous result
(v) Follow the logical direction
(vi) Relative measure, thus applicable to any field of observation, regardless of units (scales of measurements) and range of values

Summary and Conclusion

Previous studies have produced comparable information for model evaluation indices (for selected models or in general). But no comprehensive standardization (or concrete suggestion) is available including recently developed indices. The purpose of this investigation is to review and evaluate available indices for model performance evaluation and explore a logical, interpretable, and unambiguous index for general use in model evaluation. The r, R^2, and RMSE have been regarded as non-logical, ambiguous and mis-interpretable from previous studies (and have been suggested to abundant from the array of performance testing indicators) and also from this study.

The present investigation demonstrates that the index of agreement (d) between very dissimilar model-predicted variable and observed data can approach to one (1.0), but can have lower value for nearly similar data sets. The ambiguous and inconsistent behavior of d_r are also observed, thus cannot be regarded as a reliable indicator. The investigation also demonstrates that the efficiency based indicators such as E_{NS} and E_{LG}, are not consistent with logical trend (and shows reverse trend in some cases), and also with widely accepted difference-based measures (e.g. MAE, RMSE).

The PMARE (which is based on similar principle of MAE, but relative to observed data) shows consistent, robust, descriptive (clear interpretative), and logical behavior, and thus can be used as an ideal indicator for model evaluation under diverse output conditions. The performance rating based on PMARE is also suggested. From investigation of various data sets (diverse in nature), it can be concluded that the index is measuring error with both accuracy and precision.

References

Addiscott T. M., Whitmore A. P.,1987. Computer simulation of changes in soil mineral nitrogen and crop nitrogen during autumn, winter and spring. J. of Agril. Sci., Cambridge, 109, 141 – 157.

Alexandrov, G.A., D. Ames, G. Bellocchi, B. Michael, N. Crout, M. Erechtchoukova, A. Hildebrandt, F. Hoffman, C. Jackisch, P. Khaiter, G. Mannina, T. Matsunaga, S.T. Purucker, M. Rivington, L. Samaniego. 2011. Technical assessment and evaluation of environmental models and software : Letter to the Editor. Environ. Modell. and Soft. 26 (3): 328-336.

Bellocchi G., Acuit M., Fila G., Donatelli M., 2002. An indicator of solar radiation model performance based on a fuzzy expert system. Agron. J. 94, 1222-1233.

Confalonieri, R., S. Bregaglio, S. Bocchi, M. Acutis. 2010. An integrated procedure to evaluate hydrological models. Hydrol. Proces. 24(19): 2762-2770

Crescimanno G., Garofalo P., 2005. Application and evaluation of the SWAP model for simulating water and solute transport in a cracking clay soil. Soil Sci. Soc. Am. J. 69, 1943-1954.

Dust M., Baran N., Errera G., Huston J.L., Mouvet C., Schafet H., Vereecken H., Walker A., 2000. Simulation of water and solute transport in field soils with the LEACHP model. Agric. Water Manage. 44, 225-245.

Fox D.G., 1981. Judging air quality model performance. Bull. Am. Meteorol. Soc. 62, 599-609.

Geerts S., Raes D., Garcia M., Miranda R., Cusicanqui J. A., Taboada C., Mendoza J., Huanca R., Mamani A., Condori O., Mamani J., Morales B., Osco V., Steduto P., 2009. Simulating Yield Response of Quinoa to Water Availability with AquaCrop. Agron. J. 101(3), 499-508.

Jacovides C.P., Kontoyiannis H., 1995. Statistical procedures for the evaluation of evapotranspiration models. Agric. Water Manage. 27, 365-371.

Legates D.R, McCabe G.J. Jr. 1999. Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic
model validation. Water Resources Research 35(1): 233–241.

Liu Y., Teixeira J.L., Zhang H.J., Pereira L.S., 1998. Model validation and crop coefficients for irrigation scheduling in the North China plain. Agric. Water Manage. 36, 233-246.

Loague K., Green R. E., 1991. Statistical and graphical methods for evaluating solute transport models: Overview and application. J. Contam. Hydrol. 7, 51 – 73.

Martorana F., Bellocchi G., 1999. A review of methodologies to evaluate agroecosystem simulation models. Ital. J. Agron. 3(1), 19-39.

McCuen, R.H., Z. Knight, A.G. Cutter. 2006 . Evaluation of the Nash–Sutcliffe Efficiency Index. J. Hydrologic Engrg. 11: 597-602.

Moriasi D.N., Arnold J.G., Van Liew M.W., Bingner R.L., Harmel R.D., Veith T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE. 50(3), 885-900.

Nash J. E., Sutcliffe J. V., 1970. River flow forecasting through conceptual models. 1. A discussion of principles. J. Hydrol. 10, 282- 290.

Prasher S.O., Madani A., Clemente R.S., Geng G.Q., Bhardwaj A., 1996. Evaluation of two water table management models for Atlantic Canada. Agric. Water Manage. 32, 49-69.

RANDOM.ORG. 2012. Random Integer Generator. www.random.org/integers, Accessed on 20th Feb. 2012.

Rivington M., Bellocchi G., Matthews K.B., Buchan K., 2005. Evaluation of three model estimations of solar radiation at 24 UK stations. Agril. and Forest Meteorl. 132, 228-243.

Shen Z.Y., Gong Y.W., Li Y.H., Hong Q., Xu L., Liu R.M., 2009. A comparison of WEPP and SWAT for modeling soil erosion of the Zhangjiachong watershed in the three Gorges reservoir area. Agric. Water Manage. 96, 1435-1442.

Steduto P., Hsiao T.C. , Raes D., Fereres E., 2009. AquaCrop—The FAO crop model to simulate yield response to water: I. Concepts and underlying principles. Agron. J. 101(3), 101- 426.

Stockle C., Martin S., Campbell G., 1992. A model to assess environmental impact of cropping systems. Am.Soc.Agric.Eng. Paper No. 92-2041.

Stockle C.O., Donatelli M., Nelson R., 2003. CropSyst – a cropping systems simulation model. Eur.J.Agron. 18, 289-307.

Suleiman A.A., 2008. Modeling daily soil water dynamics during vertical drainage using the incoming flow concept. Catena 73, 312-320.

Wagener T., Kollat J., 2007. Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox. Environ. Modelling & Softwares, 22, 1021-1033.

Willmott C. J., 1981. On the validation of models. Phys. Geogr. 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., Ackleson S.G., Davis R.E., Feddeema J.J., Klink K.M., Legates D.R., O'Connell J., Rowe C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90(C5), 8995-9005.

Willmott, C.J., and K. Matsuura. 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolates. Int. J. Geogr. Infor. Sci. 20(1): 89-102

Willmott,C. J., Robeson, S. M. and Matsuura, K. (2011). A refined index of model performance. Int. J. Climatol. DOI: 10.1002/joc.2419

Yang C-C., Prasher S.O., Wang S., Kim S.H., Tan C.S., Drury C., Patel R.M., 2007. Simulation of nitrate-N movement in southern Ontario, Canada with DRAINMOD-N. Agric. Water Manage. 87, 299-306.

Yang, J., Greenwood D.J., Rowell D.L., Wadsworth G.A., Burns I.G., 2000. Statistical methods for evaluating a crop nitrogen simulation model, N_ABLE. Agric. System 64:37-53