Agricultural Systems 127 (2014) 81–89

Contents lists available at ScienceDirect

Agricultural Systems

journal homepage: www.elsevier.com/locate/agsy

An evaluation of the statistical methods for testing the performance of crop models with observed data

J.M. Yang [a], J.Y. Yang [b,*], S. Liu [b,c], G. Hoogenboom [d]

[a] College of Resource & Environment Sciences, Jilin Agricultural University, Changchun 130118, PR China
[b] Greenhouse and Processing Crops Research Centre, Agriculture & Agri-Food Canada, Ontario N0R 1G0, Canada
[c] Shanxi University, Institute of Loess Plateau, Wucheng Road 92, Xiaodian District, 030006 Taiyuan, PR China
[d] AgWeatherNet, Washington State University, WA 99350-8694, USA

ARTICLE INFO

Article history:
Received 19 July 2013
Received in revised form 1 January 2014
Accepted 5 January 2014
Available online 28 February 2014

Keywords:
Statistical evaluation
Test statistics
Deviation statistics
Autocorrelation
Heteroskedasticity
Crop simulation model

ABSTRACT

Calibration and evaluation are two important steps prior to the application of a crop simulation model. The objective of this paper was to review common statistical methods that are being used for crop model calibration and evaluation. A group of deviation statistics were reviewed, including root mean squired error (*RMSE*), normalize-*RMSE* (*nRMSE*), mean absolute error (*MAE*), mean error (*E*), paired-*t*, index of agreement (*d*), modified index of agreement (d_1), revised index of agreement (d_1'), modeling efficiency (*EF*) and revised modeling efficiency (EF_1). A case study of the statistical evaluation was conducted for the DSSAT Cropping System Model (CSM) using 10 experimental datasets for maize, peanut, soybean, wheat and potato from Brazil, China, Ghana, and the USA. The results indicated that R^2 was not a good statistic for model evaluation because it is insensitive to regression coefficients (α and β) of the linear model $y = \alpha + \beta x + \varepsilon$. However, linear regression can be used for model evaluation (test H0: $\alpha = 0$, $\beta = 1$) if auto-correlation, normality and heteroskedasticaity of the error term (ε) are tested or the proper data transfers are made. The results also illustrated that statistical evaluation of total dataset across treatments might be insufficient. Hence the evaluation of each treatment is necessary to make the right conclusion, especially when evaluating soil water content under different planting date treatments and soil mineral N under different N treatments. Co-variability analysis among dimensionless statistics (d, d_1, d_1', *EF* and EF_1) recommended that *d* and *EF* are inflated by the sum of squares-based deviations, i.e., the larger deviations contribute more weight on the statistic than the smaller deviation due to the squared term. However, *EF* had a larger range and a clear physical meaning at $EF = 0$, making it superior to *d*. Values of $d = 0.75$ were obtained from regression with all positive values of *EF* ($EF \geqslant 0$), indicating that values of $d \geqslant 0.75$ and $EF \geqslant 0$ should be the minimum values for plant growth evaluation. Values of $d \geqslant 0.60$ and $EF \geqslant -1.0$ should be the minimum values for soil outputs evaluation combined with *t*-test due to the fact that the soil parameters in the DSSAT SOIL module are difficult to calibrate compared with plant growth parameters because of no sufficient observed soil dataset. Due to the statistical nature, no single statistic is more robust over others but some statistics are highly correlated. Therefore, several statistics may be used from each of the following correlated groups (*RMSE*, *MAE*), (*E*, *t*-test), (d, d_1, d_1') and (*EF*, EF_1) in one assessment of model evaluation so that a representative statistical conclusion can be obtained with respect to model performance.

1. Introduction

Crop simulation models are simplified representations of real crop growth processes and the empirical equations that normally describe these dynamic processes are based on a certain set of hypotheses, which result in simulation bias or errors. Therefore, the evaluation of model performance with the observed data is important and a statistical evaluation is regarded as the key method in comparing model outputs with the observed data (Willmott, 1982; Willmott et al., 1985; Reckhow et al., 1990; Yang et al., 2000). Classical linear regression $y = \alpha + \beta x + \varepsilon$ is often used for model evaluation, and R^2 is used to assess the quality of fit of the linear model. However, the error term, ε, in the ordinary least square linear regression is based on three assumptions, namely, independence, normality and homoscedasticity. Many observed/

* Corresponding author. Tel.: +1 519 7381270; fax: +1 519 7382929.
E-mail address: Jingyi.yang@agr.gc.ca (J.Y. Yang).

http://dx.doi.org/10.1016/j.agsy.2014.01.008
0308-521X/© 2014 Elsevier Ltd. All rights reserved.

measured datasets violate some of these assumptions, making the statistical analysis inaccurate if no proper data transformations are applied (Aigner, 1971; Snedecor and Cochran, 1976; Yang et al., 2000).

Moreover, the R^2 is mainly a measure of the correlation between y and x and it is insensitive to additive (regression intercept) and proportional differences (regression slope) between the simulated and observed data (Willmott et al., 1985; Legates and McCabe, 1999). It can easily obtain a R^2 value close to 1.0 even if a model systematically over- or underestimates the observed data (Krause et al., 2005). Some deviation statistics were established to test deviation $y - x$ directly, where y and x represent the simulated and observed data, respectively (Willmott, 1981; Willmott et al., 1985). Reckhow et al. (1990) compared test statistics with model evaluation and Smith et al. (1997) applied both test statistics (linear regression and t-tests) and several deviation statistics (root mean square error, modeling efficiency and index of agreement etc.) for evaluating nine soil models using seven long-term field experiments because each statistical method provided information on a distinct aspect of the accuracy of the simulation. Yang et al. (2000) systematically tested normality and heteroskedasiticity of the error term and concluded that data from field experiments easily violated normality and equal variance assumptions, and data transformations were suggested for regression analysis between the simulated and the observed data. Sinclaira and Seligman (2000) suggested that the evaluation of the model should not only be based on end-of-season data, but also data observed during the entire growth cycle. Recent reviews indicated that nearly all ecological data are autocorrelated in both space and time (Boyce et al., 2010). More examples of autocorrelation can be seen from biomass growth of forest trees (Brienen et al., 2006) and soil evapotranspiration (Medeiros et al., 2012). Rykiel (1996) provided a comprehensive review on model validation, criteria and steps.

The Decision Support System for Agrotechnology Transfer (DSSAT) is a software package that encompasses the Cropping System Model (CSM) for over 25 different crops. DSSAT has been widely used by over 2000 scientists and extension experts globally (Tsuji et al., 1994; Hoogenboom et al., 2010; Jones et al., 2003), However, statistical methods for calibration and evaluation of the DSSAT CSM were not reported systematically in the above published studies although concepts for calibrating and evaluation of crop growth models were introduced in DSSAT v3 volume 4 (Hoogenboom et al., 1999). The objective of this paper was to review statistical evaluation methods with case evaluation examples of the DSSAT v4.5 model, and to discuss the advantages and limitations of these statistical methods which are generally applicable for the evaluation of other crop, soil and hydrologic models.

2. Reviews of evaluation statistics

2.1. Test statistics

2.1.1. Linear regression and R^2

The correlation–regression based statistical methods, such as a linear regression, correlation coefficient (r) and coefficient of determination (R^2) have frequently been used to explain how well the simulated data, y (dependent variable), against the observed data, x (independent variable). After a statistical test on regression (F or R^2), further statistical tests of the null hypothesis H_0: $\alpha = 0$, $\beta = 1$ should be carried out (Willmott, 1981; Kobayashi and Salam, 2000; Yang et al., 2000; Moriasi et al., 2007). The linear model is;

$$y = \alpha + \beta x + \varepsilon \tag{1}$$

where α and β are the regression intercept and slope respectively, and ε is a random error. The t-test can be used to test the H0: $\alpha = 0$ and H0: $\beta = 1$ because $t_a = (a - \alpha)/S_a$ and $t_b = (b - \beta)/S_b$ follow a t (n–2) distribution, where a and b are the ordinary least square measures of the parameter α and β in Eq. (1).

H_0 is maintained when the intercept (α) and slope (β) are not significantly different from 0 and 1, respectively'. If the variance of error, ε, is larger, the differences between y and x will be important even if the regression parameter, β, is not significantly different from one. In this case, regression significance should be tested by either F test or R^2. When testing R^2, H_0 hypothesis should be $\beta = 0$ because the R^2 is insensitive to the intercept α, Particularly, the R^2 measures the proportion of the variation in y which is accounted by the linear model $y = \alpha + \beta x + \varepsilon$. Thus, R^2 tests the "goodness of fit" of the linear model Eq. (1) with clear physical meaning of the R^2 value between 0 and 1. The $R^2 = 1$ indicates a perfect fit of Eq. (1), and the $R^2 = 0$ indicates that there is no linear relation.

Unfortunately, the R^2 has a serious limitation on the evaluation of model outputs that has been reported since the 1970s (McCuen and Snyder, 1975; Willmott, 1981; Legates and McCabe, 1999; Kobayashi and Salam, 2000). Because $R^2 = r^2$, we only focus on R^2 in following discussion. In fact, R^2 only estimates linear relationship between two variables and it is not sensitive to additive (regression intercept a) and proportional differences (regression slope b) between the model simulated and observed data (Willmott, 1981). In other word, $R^2 = 1$ can be demonstrated with any non-zero value of intercept a slope b in Eq. (1) (Legates and McCabe, 1999). To avoid this, Krause et al. (2005) provided a weighted version of $wR^2 = |b| \times R^2$ $(b \leqslant 0)$ and $wR^2 = b^{-1} \times R^2$ $(b > 0)$ under condition of coefficient $a = 0$. Moriasi et al. (2007) provided model evaluation guidelines and also addressed the limitation of R^2 for the model evaluation although both r and R^2 are still being used for model evaluations (Akinremi et al., 2005; Rinaldi et al., 2007; Cao et al., 2012).

2.1.2. Student t test

In model evaluation, a student t test can be used to test the zero hypothesis H_0: $\bar{d} = (\mu_y - \mu_x) = 0$, where μ_y and μ_x are the population means of the simulated data y_i and observed data x_{i}, $i = 1,2,\ldots,n$, respectively,

$$Paired - t = \bar{d}/s\bar{d} \tag{2}$$

where $S_{\bar{d}} = \sqrt{\sum (d_i - \bar{d})^2 / n(n-1)}$ is the standard deviation of $\bar{d}$. The student t test has a different purpose. In the model evaluation, it was used to test whether the model mean deviation ($\bar{d}$) differ significantly from 0. If the calculated $|t| \leqslant t_\alpha\,(n-1)$, where $t_a(n-1)$ is a critical value assuming that the true mean difference is zero, $\alpha = 0.05$ (or 0.01) is the significant levels with $n - 1$ the degree of the freedom, we conclude that the model simulated average y has no statistical difference from the observed average x. It is true that if $y_i = x_i$ for all $i = 1,\ldots,n$, then $\bar{y} = \bar{x}$, but the converse clearly is not true; meaning that $\vec{y} = \bar{x}$ does not imply $y_i = x_i$ for all simulated points n, Therefore, the t-test is just a test of unbiasedness of the mean values.

In the ordinary least square linear model (Eq. (1)), the error term ε is assumed to be normal, independent and having a uniform variance and in the student-t test (Eq. (2)), $\bar{d}$ is assumed a normal variable. In crop growth and soil process simulation, time series variables (such as aboveground biomass, soil water content and soil N dynamics) are accumulated data over time (i.e., days from planting to harvest), meaning that these data may have autocorrelation with values in pasts periods and this will result in the error term, ε, in Eq. (1) to exhibit autocorrelation and/or heteroskedasticaity. For this reason, tests of autocorrelation and heteroskedasticity are necessary before using the least square linear regression. Test results revealed that most of the time series datasets showed significant autocorrelation and heteroskedasticity and almost half of the datasets violated the normal distribution in the error term (please read Appendix A for details).

2.2. Deviation statistics

To overcome the limitation of correlation-based statistics, efficiency measures have been developed in recent decades to test deviation ($d = y - x$) directly. A simple statistical index is the mean error (E) (Addiscott and Whitmore, 1987; Yang et al., 2000)

$$E = \sum (y_i - x_i)/n \tag{3}$$

where $i = 1,2,\ldots,n$. The mean error E measures whether the model simulated y tend to overestimate ($E > 0$) or underestimate ($E < 0$) the observed x data. E is identical to $\bar{d}$ in Eq. (2), Therefore, the significant difference of E can be tested by a paired-t test (Akinremi et al., 2005; Liu et al., 2013). The disadvantage is that the positive and negative errors can negate each other in the E value. For instance, a larger positive and negative deviations can still yield a $E = 0$. Due to this limitation, the sum of squares-based measures was developed. This paper only discusses root mean square error (*RMSE*), modeling efficiency (*EF*) and index of agreement (d) as below:

$$RMSE = \sqrt{\sum (y_i - x_i)^2/n} \tag{4}$$

The relative *RMSE* is expressed as

$$nRMSE = RMSE/\bar{x} \times 100 \tag{5}$$

$$EF = 1 - \sum (y_i - x_i)^2 / \sum (x_i - \bar{x})^2 \tag{6}$$

$$d = 1 - \sum (y_i - x_i)^2 / \sum (|y_i - \bar{x}| + |x_i - \bar{y}|)^2 \tag{7}$$

The *RMSE* take on the same unit of deviation, $y - x$, and it is commonly used in both model calibration and validation processes (Loague and Green, 1991). In testing, it was found that the *RMSE* could not be used for inter comparisons for many state variables with different units, such as biomass, LAI and N concentration in plant. In these cases, *nRMSE* is used as a relative measure for inter comparisons of different variables or different models (Priesack et al., 2006).

The modeling efficiency, *EF* ($-\infty$ to 1), was first introduced by Nash and Sutcliffe (1970) to test modeling efficiency of their river flow model and later was widely used in modeling evaluation with several different names; such as modeling efficiency (Loague and Green, 1991; Mayer and Butler, 1993); Nash–Sutcliffe Efficiency (NSE) (Moriasi et al., 2007) and coefficient of efficiency (Legates and McCabe, 1999; Willmott et al., 1985, 2011). *EF* is a dimensionless measure, an $EF = 1$ corresponds to a perfect match of modeled output with the observed data and $EF < 1$ for any realistic simulation. $EF < 0$ if the model predicted values are worse than simply using the observed mean ($\bar{x}$) to replace the simulated y_i. To some extent, $EF \geqslant 0$ is a critical condition to conclude "goodness of match" between the simulated and the observed.

The index of agreement, d ($0 \leqslant d \leqslant 1$), is a dimensionless and bounded measure originally provided by Willmott (1982). It has been recommended by many modelers to conduct cross-comparisons between simulated and observed data (Legates and McCabe 1999; Krause et al., 2005; Moriasi et al., 2007). Similar to *EF*, d statistic is a sum of squares-based, dimensionless statistics, mainly used to depict the degree to which the deviation toward zero. Both statistics are suitable to compare accuracies of many output variables together.

However, the main disadvantages of the sum of squares-based statistics (such as *EF* and d) are that they are more sensitive to larger deviations than smaller deviations. Legates and McCable (1999) pointed out that *EF* was overly sensitive to extreme values, as was R^2 because of the squared differences. Krause et al. (2005) summarized that the largest disadvantage of *EF* and d statistics was the fact that the differences between the simulated and observed values were calculated as squared values. As a result, these sums of squares-based statistics are overly sensitive to outliers or larger deviations due to the squaring of the deviation term (Willmott et al., 2011). This can result in that the *EF* and d are strongly influenced by larger (higher) deviations while smaller deviations are counted much less in a time series (Legates and McCabe, 1999; Krause et al., 2005).

As discussed above, the sum of squares-based statistics are easily inflated by the squaring the deviation term, the sum of absolute values-based statistics were therefore constructed to overcome this problem (Willmott et al., 1985, 2011; Legates and Mccabe, 1999). The mean absolute error (*MAE*), the modified modeling efficiency (EF_1), the modified index of agreement d_1 and the refined index of agreement d'_1 are all the sum of absolute values-based statistics, defined as below:

$$MAE = \sum |y_i - x_i|/n \tag{8}$$

The relative *MAE* is expressed as

$$C = MAE/\bar{x} \tag{9}$$

$$EF_1 = 1 - \sum |y_i - x_i| / \sum |x_i - \bar{x}| \tag{10}$$

$$d_1 = 1 - \sum |y_i - x_i| / \sum (|y_i - \bar{x}| + |x_i - \bar{x}|) \tag{11}$$

$$d'_1 = 1 - \sum |y_i - x_i| / 2 \sum |x_i - \bar{x}| \tag{12}$$

By definition, *MAE* reduces the inflation of the outlier compared with *RMSE* (Willmott et al., 1985) while it takes on the same units of $y - x$ and are mainly used as measures of accuracy to compare the output of the same variables. C value is a relative average statistic of the *MAE*, which is expressed as a proportion of the mean of the observed variable y. Both *MAE* and $RMSE \geqslant 0$, with the $MAE \leqslant RMSE$. The relative criteria C and *nRMSE* are correlated.

EF_1 is the modified *EF* by replacing sum of squares term by the sum of absolute values of $y - x$. Compared with *EF*, EF_1 also ranged from $-\infty$ to 1.0 but it is less sensitive to extreme values (Legates and McCabe, 1999).

The modified index of agreement, d_1, was first given by Willmott et al. (1985) to overcome the inflating effect of the extreme values in the sum of squares-based index of agreement, d. Another advantage of d_1 over d was that it approached 1.0 more slowly as simulated y approached the observed x, but it was still suboptimal due to the same narrow range of 0–1.0 and difficulty in interpreting the values (i.e., difficulty to resolve adequately the great variety of ways that simulated y can differ from the observed x)(Willmott et al., 2011). Another fact is that the d and d_1 values are relative high even though the substantial deviation is evident (see case study section). For the above reason, a refined index of agreement, d'_1, was constructed to enlarge the range from -1.0 to 1.0 compared with the range of 0–1.0 in d and d_1. d'_1 also has clear physical meanings of the value. For example, $d'_1 = 0.5$ indicates that sum of the errors magnitude is half of the sum of the perfect-model-deviation and observed-deviation magnitude Willmott et al. (2011).

In addition to EF_1, d_1 and d'_1, there are more generic forms of absolute values-based measures, by adding a power j to the absolute term $|y - x|^j$ to Eqs. (10)–(12). $j = 1$ was used in this study because they are often used and discussed in literatures (NLegates and McCabe, 1999; Krause et al., 2005; Willmott et al., 2011).

3. Case study: evaluation of DSSAT model

3.1. Methods

A case study was conducted to illustrate the evaluation statistics discussed in the previous section in relation to the DSSAT Cropping Systems Model (CSM). Four different crop modules were used in this study corresponding to the CSM-CERES-Maize, CSM-CERES-Wheat, CSM-CROPGRO-Soybean and -Peanut and CSM-SUBTOR-Potato in the DSSAT (Hoogenboom et al., 2010). The EasyGrapher software was used to conduct the statistical evaluations of the DSSAT CSM models (Yang and Huffman, 2003, 2004; Yang et al., 2010). The EasyGrapher software was developed to provide graphical and statistical analysis for DSSAT v4.x (i.e., v4.0, v4.5).

3.2. Field experiments

Ten field experiments were selected from Brazil, China, Ghana, and the USA to conduct the statistical evaluation using Eqs. (1)–(12) and to determine the co-variability of deviation statistics (EF, EF_1, d, d_1 *and* d'_1). All experiments contained about 2–10 treatments with single factor, 2 × 2 or 2 × 3 factorial designs, including one peanut experiment, five maize experiments, two soybean experiments, one wheat experiment and one potato experiment (Table 1).

One soybean experiment with three treatments of irrigation in 1981 (UFGA8101 in Table 1) and one maize experiment with 6 treatments of 3 levels of irrigation by 2 levels of fertilizer N in 1982 (UFGA8201 in Table 1) were selected from the DSSAT experiments as an "example 1" to illustrate statistical behaviors of deviation statistics, linear regression and R^2. These experiments included detailed observed plant growth dataset for model evaluation. Two maize fertilizer experiments that were conducted in 2008 and 2009 (CHUN0801 and CHUN0901 in Table 1) consisted of three nitrogen fertilizer treatments (0, 120 and 240 kg N ha^{-1}) with three replicates in Jilin China. The aboveground biomass, plant N and soil mineral N (NO_3–N and NH_4–N) contents were measured in every 2–3 weeks during the growing seasons. Details of the field experiments were reported by Yang et al. (2011a,b). The 2008 experiment was used as an "example 2" to show soil NO_3–N evaluation. One peanut experiment that was conducted in 1998 in Ghana was selected as an "example 3" to show soil water content evaluation. Six treatments (3 planting dates by 2 cultivars) were selected from this experiment with the measured soil water contents during growing season in Tamala, Ghana (see Table 1, GHNY9801) for this study. The volumetric soil water content at the different depth was measured every 5–15 days and only 5–15 cm depth data was discussed in this paper.

The time-series datasets from the above 10 experiments had 12–147 separate observations, which depended on the frequency of the measurements for each experiment during the growing season (Table 1). The observations in the time-series data included top weight, N uptake, grain yield and the maximum LAI from the USA and Brazil experiments (Table 1). The datasets are valuable for investigating the model evaluation statistics.

3.3. Statistical evaluation

3.3.1. Example 1: evaluation of plant growth and nitrogen

Evaluation example 1 was carried out using one soybean experiment and one maize experiment from the USA (see UFGA8101 and UFGA8201 in Table 1) as shown in Fig. 1. In the irrigated field in example 1, the simulated soybean top weight in the irrigated field was well matched by the observed data (Fig. 1a). The small deviation of $RMSE = 518$ kg ha^{-1} ($nRMSE = 12\%$) and $MAE = 411$ kg ha^{-1} ($C = 0.10$) were found between the simulated top weight with the observed data and a negative mean error $E = -16$ kg ha^{-1}, indicating that the model slightly underestimated the observed top weight. In the rainfed condition, the top weight was systematically underestimated by the DSSAT model compared to the observed values during growing season (Fig. 1b). The deviation was larger and effectively estimated by the $RMSE = 887$ kg ha^{-1} ($nRMSE = 21\%$), $MAE = 699$ kg ha^{-1} ($C = 0.16$). The $E = -699$ kg ha^{-1}, where the absolute value of $E = MAE$ in this example due to all negative deviations (i.e., $y_i - x_i < 0$ for all i) (Table 2). The above examples illustrate that the E value changed effectively to measures mean differences and directions (i.e., negative E values indicate underestimation). Meanwhile, the $RMSE$ ($nRSME$) and MAE (C) changed substantially from smaller values in the irrigated field in example 1 (Fig. 1a) to larger values in the rainfed condition (Fig. 1b).

Both maize tops N and stem N from example 1 were all systematically overestimated by the DSSAT model compared with the observed data (Fig. 1c and d) as indicated by positive mean error $E = 36$ kg N ha^{-1} for maize tops N, and $E = 14$ kg N ha^{-1} for maize stem N. The larger deviations between the simulated and observed tops N and stem N were estimated by values of $nRMSE = 67\%$ and 103% and $C = 0.59$ and 0.80 for maize tops N and stem N, respectively (Table 2).

The dimensionless statistics of d and EF families showed different sensitivity to deviations in both soybean and maize in example 1 simulations. Among them, d statistic is the most insensitive measure to the deviation ranging from 0.99 to 0.79 compared with the

Table 1
Experimental datasets for evaluation of the Cropping System Model for maize, peanut, soybean, wheat and potato.

Crop	DSSAT crop file	Experiment (treatment)	State or province / country	LAT	LONG	Simulated output variables	Number of observed dataset	Reference
1. Maize	CHUN0801	Notrogen (3)	Jilin China	43.9N	125.2E	Soil NH_4–N, NO_3–N	33	Yang et al. (2011a)
2. Maize	CHUN0901	Notrogen (3)	Jilin China	43.9N	125.2E	Soil NH_4–N, NO_3–N	21	Yang et al. (2011b)
3. Peanut	GHNY9801	Planting date × Cultivar (4×2)	Tamale Ghana	9.42N	−0.92W	Soil water	143–147	DSSAT v4.5
4. Maize	BRPI0202	Soil, irrigation & N (2×2×2)	SP Brazil	22.43S	47.25W	Top (leaf) weight, LAI	32–48	Soler (2004, 2005, 2007a,b)
5. Maize	IBWA8301	Cultivar & N (2×3)	Hawaii USA	21.00N	158.00W	Top (leaf, steam, grain) weight, LAI	21–33	Ritchie et al. (1993)
6. Maize	UFGA8201	Irrigation & N (3×2)	Florida USA	29.63N	82.37W	Plant N	18–66	Tsuji et al. (1998)
7. Soybean	UFGA7801	Irrigation (2)	Florida USA	29.63N	82.37W	Top (grain) weight, maximum LAI	12–28	DSSAT v4.5
8. Soybean	UFGA8101	Iirrigation (3)	Florida USA	29.63N	82.37W	Top (grain) weight, maximum LAI	30–91	DSSAT v4.5
9. Wheat	KSAS8101	Irrigation & N (2×3)	Kansas USA	37.2N	99.8W	Top (leaf, steam, grain) weight, LAI	12–72	Wagger (1983)
10. Potato	OSBO8801	Soil (10)	Oregon USA	45.8N	119.3W	Top (tuber, leaf) weight, LAI	57–109	DSSAT v4.5

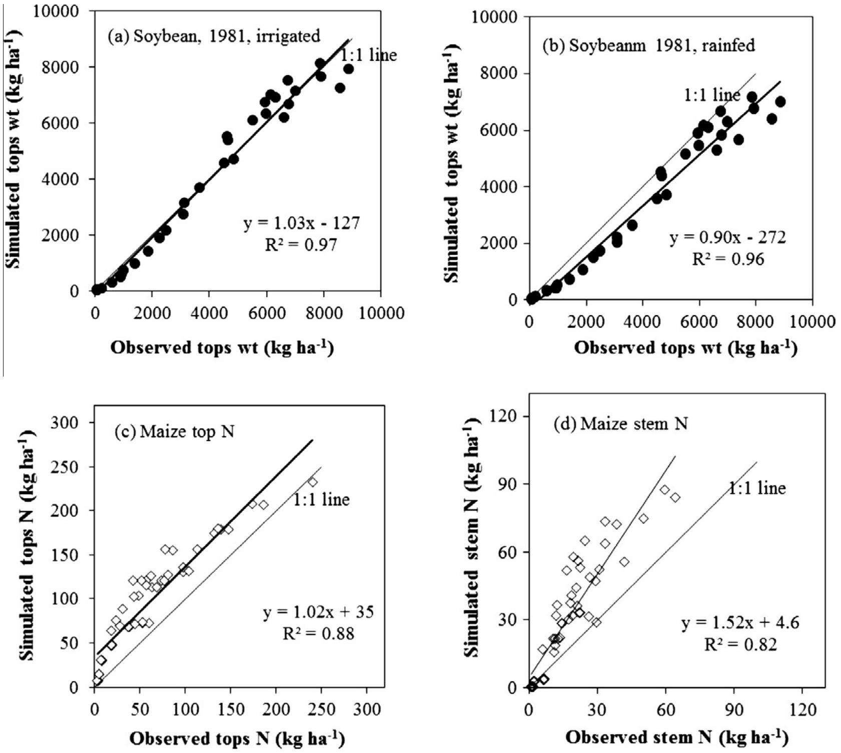

Fig. 1. Evaluation example 1: Comparison of simulated and observed above ground biomass for soybean under irrigated (a) and rainfed conditions (b) using a soybean irrigation experiment in 1981; and comparison of simulated and observed above ground N (c) and stem N (d) for maize using a maize N and irrigation experiment in 1982.

ranges of d_1 from 0.85 to 0.43 and d_1' from 0.91 to 0.33. The ranges of *EF* and EF_1 were 0.96 to −0.69 and 0.83 to −0.34, respectively among 4 output variables in example 1 (Table 2), showing that the *EF* and EF_1 had relatively larger range than *d*, d_1 and d_1'.

In evaluation of soybean and maize growth variables in example 1, linear regression and R^2 were not suggested to use due to the fact that R^2 is not sensitive to additive and proportional differences as discussed in "Testing statistics" section. To illustrate this, two linear equations were obtained for soybean top weight in both irrigated and rainfed fields between the simulated and observed values, but two R^2 values were very similar, i.e., $R^2 = 0.97$ ($y = -127 + 1.03x$) in the irrigated field, and $R^2 = 0.96$ ($y = -272 + 0.9x$) in the rainfed field (Fig. 1a and b). Similarly, in maize study in example 1, two different linear regressions were obtained but with very similar $R^2 = 0.88$ ($y = 35 + 1.02x$) for maize tops N (Fig. 1c) and $R^2 = 0.83$ ($y = 4.6 + 1.52x$) for maize stem N even if the intercept *a* and regression slope *b* were significantly different from 0 and 1 (Fig. 1d). This example clearly demonstrated that the R^2 is insensitive to the additive (estimated coefficient *a*) and proportional changes (estimated coefficient *b*). However, paired-*t* statistic can be used to test the mean difference effectively; i.e., paired-$t = -0.17$ (soybean irrigated field) and −7.12 (soybean rainfed field) and paired-$t = 12.29$ (maize tops N) and 7.65 (maize stem N) in example 1 (Table 2), indicating that only insignificant difference was found in soybean tops weight in the irrigated field while all others showed significant mean differences between the simulated and measure dataset (Fig. 1 and Table 2).

3.3.2. Example 2: evaluation of soil nitrate nitrogen

Evaluation example 2 was carried out using soil nitrate nitrogen (NO_3–N) dataset including N0, N120 and N240 treatments from the experiment 1 (CHUN0801) (Table 1). Overall statistical evaluation showed that the model underestimated the soil NO_3–N in the 0–30 cm soil layer across three treatments (Table 2). The *RMSE* (*nRMSE*) and *MAE* (*C*) values were 20.0 kg N ha^1 (81%) and 16.3 kg N ha^{-1} ($C = 0.66$), respectively, indicating a larger deviation between the simulated and measured soil NO_3–N. Further statistical evaluation by $E = -11.6$ kg N ha^{-1} and paired-$t = -4.21$ ($p > 0.01$) showed significant difference between the simulated and measured soil NO_3–N content. The calculated values of $EF = 0.10$ and $EF_1 = -0.08$ indicated that the estimated value was slightly (or no) better than the observed mean although the values of *d*, d_1 and d_1' were relative larger, i.e., 0.76, 0.48 and 0.46, respectively (Table 2).

In order to find which treatment had larger simulation error, the statistical evaluations were made on each treatment. The calculated values of *RMSE* (*nRSME*) were 18.0 (113%), 24.3 (84%), 17.2 (59%) and *MAE* (*C*) were 14.8 (0.93), 20.2 (0.70), 13.9 (0.48) kg N ha^{-1} from N0, N120 and N240, respectively, indicating that larger deviations were found in the N0 and N120 than N240 treatments. Calculated values of *E* were −14.8, −20.2 and 0.10 kg N ha^{-1} with the paired-*t* values of −4.80 ($p(t \leqslant t_a) < 0.001$), −4.96 ($p(t \leqslant t_a) < 0.001$) and 0.01 in the N0, N120 and N240 treatment, respectively, indicating that the model underestimated soil NO_3–N significantly in the N0 and N120 treatments, but no statistical mean difference was found in the N240 treatment (Table 2; Fig. 2). For comparison purpose, the R^2 (0.62) showed a high correlation between the simulated and measured soil NO_3–N in the N240 treatment while R^2 (0.04) showed the worst correlation in the N0 compared with N120 and N240 treatments (Table 2).

The calculated values of *d*, d_1 and d_1' ranged from 0.40 to 0.88, 0.33 to 0.60 and 0.00 to 0.67 in the N0, N120 and N240, respectively. The calculated values of *EF* were −2.13, −0.56, 0.60, and EF_1 were −1.00, −0.41 and 0.34, in the N0, N120 and N240 treatments, respectively, clearly indicating that poor agreements for the N0 and N120 treatments, but a reasonable agreement was in the N240 treatment (Table 2; Fig. 2). This example illustrates that the top soil NO_3–N data ranged greatly from 0.0 to 100 kg N ha^{-1} depending on treatments (i.e., fertilizer N application rates). The overall evaluation among three treatments was ineffective and it

Table 2
Statistical evaluation of the DSSAT Ccopping System Model using field experimental data from the USA, China and Ghana.

Example	Example 1				Example 2				Example 3			
Experiments	Soybean (UFGA8101)		Maize (UFGA8201)		Maize (CHUN0801)				Example 3 (GHNY9801)			
Variables	Top weight	Top weight	Top N	Stem N	Soil NO_3–N in the 0–30 cm (kg N ha^{-1})				Soil water content in the 5–15 cm (cm^3 cm^{-3})			
Treatments	Irrigated	Rainfed	All	All	Total	N0	N120	N240	Total	P170	P200	P230
Measured mean	4210	4309	62	19	24.7	15.9	29.0	29.3	0.19	0.18	0.21	0.17
Sample number	31	32	48	48	36	12	12	12	112	46	42	24
RMSE	518	887	42	19	20.0	18.0	24.2	17.2	0.06	0.07	0.04	0.04
nRMSE	12	21	67	103	81	113	84	59	29.9	40.6	20.7	21.4
MAE	411	699	37	15	16.3	14.8	20.2	13.9	0.04	0.06	0.03	0.03
C	0.10	0.16	0.59	0.80	0.66	0.93	0.7	0.48	0.24	0.33	0.16	0.19
E	−16	−699	36	14	−11.60	−14.80	−20.20	0.10	0.01	0.03	−0.01	−0.01
Paired-t	−0.17	−7.12	12.29	7.65	−4.21	−4.80	−4.96	0.01	1.51	3.43	−1.14	−1.79
$p(t \leqslant t_a)$ *two tails*	0.87	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.99	0.13	<0.001	0.26	0.09
EF	0.96	0.89	0.39	−0.69	0.10	−2.13	−0.56	0.60	0.39	−0.61	0.59	0.85
EF_1	0.83	0.71	0.10	−0.34	−0.08	−1.00	−0.41	0.34	0.21	−0.28	0.32	0.59
d	0.99	0.97	0.87	0.79	0.76	0.40	0.57	0.88	0.78	0.45	0.86	0.95
d_1	0.85	0.78	0.53	0.43	0.48	0.33	0.42	0.60	0.56	0.44	0.60	0.71
d'_1	0.91	0.86	0.55	0.33	0.46	0.00	0.30	0.67	0.61	0.36	0.66	0.80
R^2	0.97	0.96	0.88	0.83	0.47	0.04	0.55	0.62	0.41	0.01	0.61	0.97

may result in a wrong conclusion that the simulated soil NO_3–N was significantly lower than the measured dataset in all treatments (i.e., paired $t = -4.21$ and $EF = 0.10$ (Table 2). In this situation, both graphical and statistical evaluations on individual treatment are necessary to conclude that the model closely simulated soil NO_3–N in higher N condition, such as in N240 (i.e., paired $t = 0.01$, $EF = 0.60$), and it significantly underestimated soil NO_3–N in lower N condition, especially in the N0 treatment (i.e., paired $t = -4.80$, $EF = -2.13$) (Table 2, Fig. 2).

3.3.3. Example 3: evaluation of soil water content

Evaluation example 3 was carried out using soil water content dataset including three treatments of planting days from the peanut experiment in Ghana (GHNY9801 in Table 1). The simulated soil water content generally matched with the measured data across three treatments of the planting date 170, 200 and 230 (Fig. 3). Across these three treatments, *RMSE* (*nRMSE*) and *MAE* (*C*) values were 0.06 (29.9%) and 0.04 (0.24) cm^3 cm^{-3}, respectively, indicating a substantial deviation between the simulated and measured data. *E* value of 0.01 cm^3 cm^{-3} was tested by paired-*t* value of 1.51, indicating that no statistical difference was found between the simulated and measured soil water content from overall treatments. The dimensionless statistics showed overall reasonable agreements with different statistical values in the order of d (0.78) > d'_1 (0.61) > d_1 (0.56) > EF (0.39) > EF_1 (0.21) (Table 2 and Fig. 3).

When evaluating soil water content in each of three treatments (i.e., planting day 170, 200 and 230) separately, the values of the *RMSE* (*nRMSE*) were 0.07, 0.04 and 0.04 cm^3 cm^{-3} (40.6%, 20.7% and 21.4%), and values of the *MAE* (*C*) were 0.06, 0.03 and 0.03 cm^3 cm^{-3} (0.33, 0.16 and 0.19). These values indicated that the differences can be effectively estimated by the relative values of *nRMSE* and *C* although the *RMSE* and *MAE* showed no or little differences among planting day 200 and 230 treatments.

The *E* values of 0.03, −0.01 and −0.01 cm^3 cm^{-3} were tested by the paired-*t* values of 3.43, −1.14 and −1.79 in the three treatments, respectively, indicating that the statistical significant difference was only found in planting day 170 (Table 2), and this was also evidenced graphically (Fig. 3).

The dimensionless values of d, d_1, d'_1 ranged from 0.45 to 0.95, 0.44 to 0.71 and 0.36 to 0.80 from planting day 170 to 230, and the values of *EF* and EF_1 ranged from −0.61 to 0.85 and −0.28 to 0.59, indicating that planting day 230 had the best match between the simulated and measured soil water content in the 5–15 cm layer among three treatments. Among 5 dimensionless statistics, *d* showed the highest value while EF_1 showed the lowest value. On the other hand, *EF* showed the largest range of 1.46 compared with the ranges of EF_1 (0.87), d (0.50), d_1 (0.27) and d'_1 (0.44) among the three treatments. For the comparison purpose, R^2 values of 0.97 and 0.01 in the planting days of 230 and 170 treatments, respectively, showing a similar trend to the *EF* values in this example. This example illustrates that the soil water content data had a narrow range of 0.0–0.5 cm^3 cm^{-3} and values of *RMSE, MAE* and *E* were all difficult to evaluate the differences among three treatments while relative measures of *nRMSE* and *C* showed substantial differences effectively among the treatments, and paired-*t* tested significance of the mean error *E* effectively.

3.4. Co-variability among statistics

Co-variability analysis among statistics was carried out using 10 field experiments (Table 1). Of these, 51 plant growth outputs were evaluated from seven experiments in the USA and Brazil, including output state variables of top weights, stem and leaf weights, grain dry weights, LAI, harvest index and leaf N concentration etc. Total of 22 soil mineral N outputs (soil NO_3–N plus NH_4–N) were

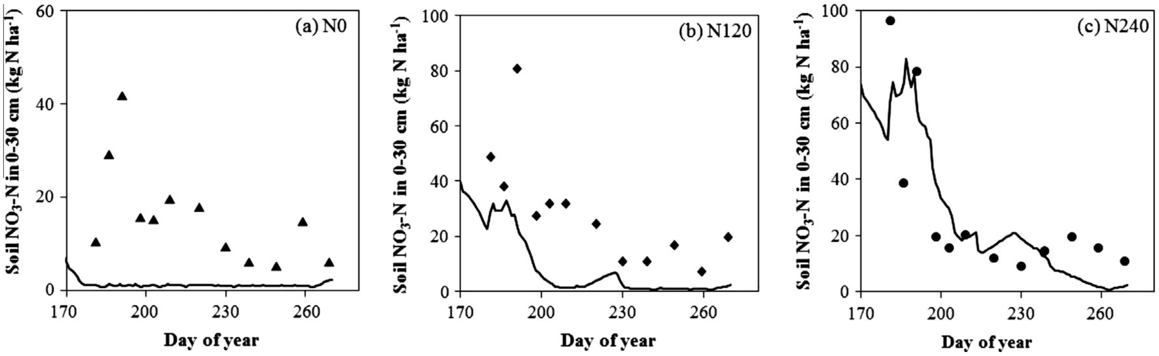

Fig. 2. Evaluation example 2: Comparison of the simulated soil NO_3–N (line) in the soil profile with the measured dataset (symbol) in the growing seasons under the N0 (a), N120 (b) and N240 (c).

evaluated using two nitrogen experiments (CHUN0801, CHUN0901) in Jilin China (Table 1), including total soil mineral N in whole soil profile, NH_4–N and mineral N contents from the 15–30, 30–45, 45–60, 60–70 and 70–80 cm layers.

From the definition, we know that the *EF* and EF_1 values ranged from $-\infty$ to 1, *d* and d_1 ranged from 0 to 1 and d_1' ranged from -1 to 1. Because the $EF = 0$ has a clearly physical meaning, we are interested in analyzing the ranges of *d*, d_1 and d_1' when *EF* (or EF_1) value approached 0. To illustrate these relationships, we need to analyze co-variability between *d*, d_1 and d_1' versus *EF* (Fig. 4a and b); *d*, d_1 and d_1' versus EF_1 (Fig. 4c and d); EF_1 versus *EF* (Fig. 4e) and *d*, d_1 versus d_1' (Fig. 4f).

It is clear that *d* differs greatly from *EF* in our example, the calculated *EF* values ranged from -7.04 to 0.98 and *d* values ranged from 0.42 to 0.99 (Fig. 4a). If we sort all statistics from the largest to the smallest by *EF* values, then regress these *d* values that corresponds to all positive *EF* values, we obtain a linear regression is $d = 0.7531 + 0.244\ EF$ with $R^2 = 0.91$. This illustrates that *EF* and *d* values can be expressed as a linear relationship when *EF* values greater than 0. When $EF = 0$, $d = 0.75$; when *EF* values < 0, *d* value showed much scatter positive values between 0 and 0.75 (Fig. 4a). Negative values of *EF* indicated that the simulation was worse than simply using the measured mean, i.e., a very poor match between the simulated and measured data. When *d* statistic is used alone, caution should be made because values of *d* can be relatively high (>0.70) despite the poor match (i.e., $EF < 0$). This disadvantage of *d* was reported by Krause et al. (2005). The ranges of d_1 and d_1' were 0.23 to 0.88, and -0.41 to 0.93, respectively, and both d_1 and d_1' showed positive correlations with *EF* (i.e., $r = 0.84$ and 0.95, respectively) (Fig. 4b). The values of $EF = 0$ corresponded the values of d_1 and d_1' at 0.50.

The range of values for EF_1 was -2.37 to 0.86 and the correlation between *d* and EF_1 was similar to *d* and *EF* (Fig. 4c). A linear regression of these *d* values that corresponds to all positive EF_1 values was $d = 0.7938 + 0.241\ EF_1$ with $R^2 = 0.90$. When EF_1 values < 0, *d* value showed much scatted positive values between 0 and 0.79. Both d_1 and d_1' showed curve linear relationships with EF_1 (Fig. 4d), and the correlation coefficients between d_1, d_1' and EF_1 were 0.95 and 1.00, respectively. The value of $EF_1 = 0$ corresponded to the values of d_1 and d_1' of 0.50.

The values of *EF* and EF_1 showed close positive correlation ($r = 0.95$) but the range of *EF* was larger than EF_1 (Fig. 4e). The values of d_1 and d_1' showed close curve linear relationship with correlation coefficient of 0.95 (Fig. 4f). The values of *d* and d_1' showed similar positive correlation pattern to *d* and EF_1 with $r = 0.83$. Compared *d*, d_1 with d_1', it was evident that the ranges of d_1' doubled that of *d*, d_1.

4. Discussion

The values of *d* showed a narrow range when using only 51 crop growth variables from seven USA and Brazil experiments (Table 1). For instance, *d* values ranged from 0.90 to 0.99 in 34 of 51 plant growth variables/outputs from seven experiments only, while d_1 and d_1' ranged from 0.57 to 0.88 and 0.62 to 0.93, respectively. This illustrated that d_1 and d_1' had better value ranges than *d*. In the same dataset, *EF* and EF_1 ranged from 0.56 to 0.98 and 0.23 to 0.86, respectively, indicating that both *EF* and EF_1 had reasonable ranges while $EF > EF_1$. For all dimensionless measures, the *EF* and d_1' showed good statistical behaviors in that they have a clearly physical meaning when the values approached zero. When using these statistics to evaluate plant growth outputs from the DSSAT CSMs, $EF > 0$ and $d \geqslant 0.75$ should be the minimum values for top (leaf, stem) weights, grain yield and LAI.

When analyzing the 30 output variables in soil water contents and soil mineral N experiments from Tamale Ghana and China's

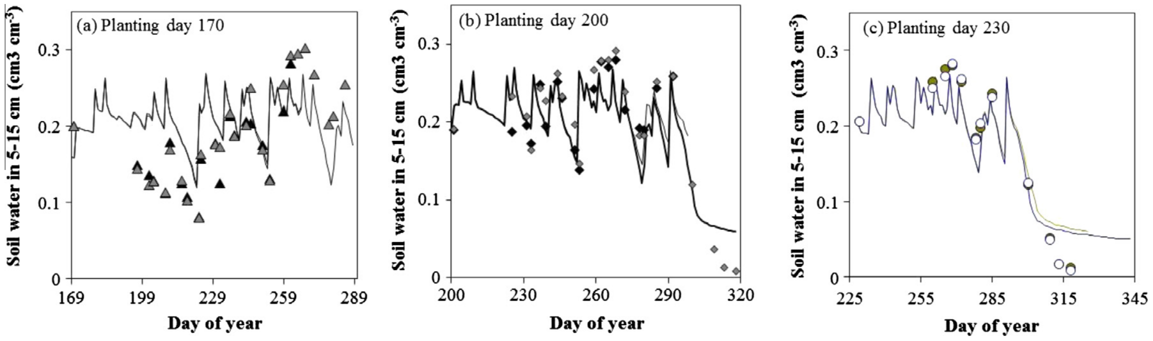

Fig. 3. Evaluation example 3: Comparison of simulated soil water content (line) with the measured dataset (symbol) for peanut under three treatments each with two cultivars in the growing season under planting days of 170 (a), 200 (b) and 230 (c).

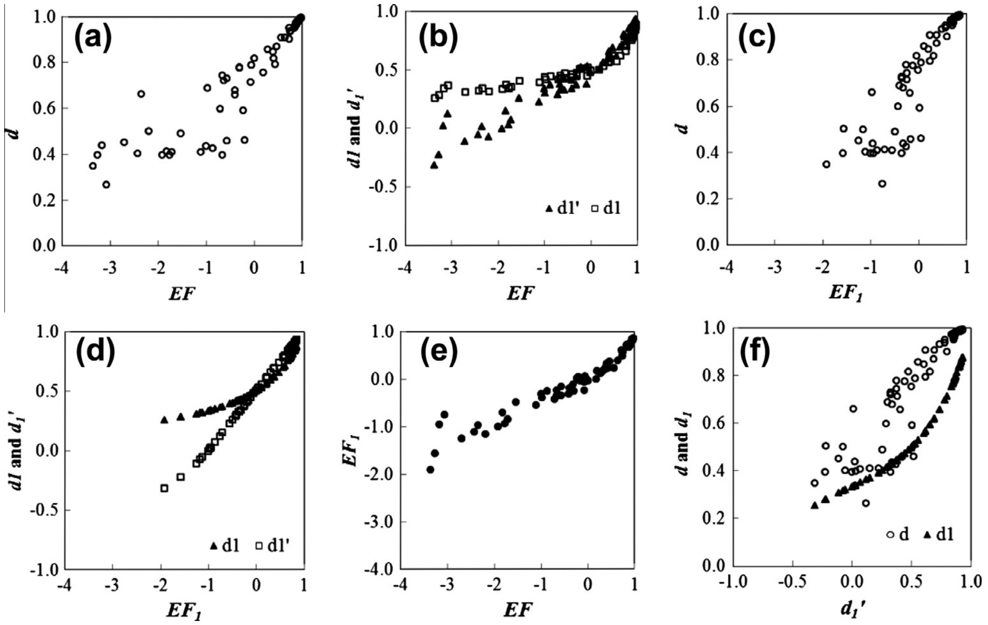

Fig. 4. Co-variability analysis of deviation statistics from 81 calculated values of d (open circle), d_1 (triangle), d'_1 (open square) and EF_1 (circle) with EF using dataset from Table 1.

experiments (Table 1), d, d_1 and d'_1 ranged from 0.57 to 0.88, 0.23 to 0.51 and −0.41 to 0.52, respectively. This illustrated that both d and d_1 had smaller ranges than d'_1. In the same 30 dataset, EF and EF_1 ranged from −0.74 to 0.18 and −2.37 to 0.00, respectively, and 8 of 30 paired-t values showed no statistical difference. This illustrated that the model simulated soil water and soil mineral N contents may have larger deviations than the plant growth variables, such as top weight and yield. The larger deviations in simulated soil water and soil mineral N contents than top weight and yield can be explained by the fact that the DSSAT model calibrations were usually made on the plant growth (i.e., top weight, yield and LAI) but few calibrations was on the soil water and soil nitrogen parameters due to the limited observation dataset.

Therefore, the evaluation results of d and EF values on soil outputs might be interpreted using a different subjective criteria from plant growth variables, such as that a reasonable agreement could be made from $EF \geqslant -1.0$ and $d \geqslant 0.60$ together with a t-test.

In addition, a statistical evaluation should be made individually from each treatment if N treatments (levels) are largely different from each other. All statistics were sensitive to the larger deviations between the simulated and measured soil mineral N. Among these, t-test is a good statistic to test significant difference of the mean error E; relative measure, $nRMSE$ and C were suggested to diagnose the average deviation. The reasons why the DSSAT soil model underestimated soil NO_3–N in the N0 and N120 treatments can be explained as follows: it is possible that the model calibration was made based on ample N condition, such as N240 in this study, and the model did not have sufficient input data of manure application or previous crop credit.

5. Recommendation

Statistical analysis is an important procedure during model calibration and evaluation, but there is no standard way on how many and which statistic should be used. Correlation based statistics (r and R^2) are not suggested for model evaluation because both are insensitive to additive (regression intercept) and proportional differences (regression slope). Lineal regression (test H0: $a = 0$ and $b = 1$) can only be used to evaluate the model outputs with the observed data when the time series datasets follow three assumptions of independence, normality and homoskedasitacity in the error term. For the paired t statistic, the difference $\bar{d}$ follows assumptions of independence and normality and it does not need the equal variance (Snedecor and Cochran, 1976; Yang et al., 2000). The deviation statistics, by definitions of Eqs. (3)–(10), do not need the error term follow three assumptions because they are not hypothesis tests. However, it was assumed that all the error in deviation statistics was contained within the simulated variable y, and that the observed x is error free (Willmott et al., 1985).

Deviation statistics $RMSE$ and MAE are good for use in model calibration stage because both have the same unit as the observed and simulated variables. Mean error E is a good index to determine if the model under- (negative) or over-estimates (positive) the observed data, while the paired-t can be used to test the significant difference of E. The dimensionless measures d, d_1 and d'_1, EF and EF_1 are all widely used deviation statistics for model evaluation. Among them, the value for d is easily inflated by the sum of squires-based deviations, and has large values even if the simulation is poor. The value of EF is also inflated by the sum of squares-based deviations, but EF has a larger range that makes it superior to d. In this paper, we suggest that the $EF \geqslant 0$ and $d \geqslant 0.75$ should be the minimum values for evaluating plant growth outputs, while $EF \geqslant -1.0$ and $d \geqslant 0.60$ should be the minimum values for evaluating soil water and mineral N outputs together with a t-test result.

Both d_1 and d'_1 and EF_1 are the sum of absolute errors-based statistics, and had larger ranges than d, but these statistics are difficult to obtain a larger value. Due to the statistical nature, no single statistic is robust over another but some statistics are highly correlated that should not be used in the same evaluation. On the other hand, statistics can be select to capture different aspects of the model and observation difference. Therefore, several statistics may be selected from each of following correlated groups ($RMSE$, MAE), (E, t-test), (d, d_1, d'_1) and (EF, EF_1) in one assessment of model evaluation so that a representative statistical conclusion can be drawn from them. It should be noted that d'_1 and EF_1 are equal in this paper and they should not be used in the same evaluation process. The statistical evaluation methods that were applied in this project for DSSAT can generally be applied for the evaluation of other simulation models.

Acknowledgements

We thank the DSSAT team for providing support for the use of experimental data that are provided with the DSSAT distribution version. DSSAT can be obtained from the DSSAT Foundation (www.DSSAT.net).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.agsy.2014.01.008.

References

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

Aigner, D.J., 1971. Basic Econometrics. Prentice-hall Inc., Englewood Cliffs, New Jersey.

Akinremi, O.O., Jame, Y.W., Campbell, C.A., Zentner, R.P., Chang, C., de Jong, R., 2005. Evaluation of LEACHMN under Dryland conditions. I. Simulation of water and solute transport. Can. J. Soil Sci. 85, 223–232.

Boyce, M.S., Pitt, J., Northrup, J.M., Morehouse, A.T., Knopff, K.H., Cristescu, B., Stenhouse, G.B., 2010. Temporal autocorrelation functions for movement rates from global positioning system radiotelemetry data. Phil. Trans. R. Soc. B 365, 2213–2219.

Brienen, R.J.W., Zuidema, P.A., During, H.J., 2006. Autocorrelated growth of tropical forest trees: unraveling patterns and quantifying consequences. Forest Ecol. Manage. 237, 179–190.

Cao, H., Hanan, J.S., Liu, Y., Liu, Y.X., Yue, Y.B., Zhu, D.W., Lu, J.F., Sun, J.Y., Shi, C.L., Ge, D.K., Wei, X.F., Yao, A.Q., Tian, P.P., Bao, T.L., 2012. Comparison of crop model validation methods. JIA 11, 1274–1285.

Hoogenboom, G., Wilkens, P.W., Tsuji, G.Y., 1999. DSSAT v3, vol. 4. University of Hawaii, Honolulu, Hawaii.

Hoogenboom, G., Jones, J.W., Wilkens, P.W., Porter, C.H., Hunt, L.A., Boote, K.J., Singh, U., Uryasev, O., Lizaso, J.I., Gijsman, A.J., White, J.W., Batchelor, W.D., Tsuji, G.Y., 2010. Decision support system for agrotechnology transfer (DSSAT) version 4.5 [CD-ROM], University of Hawaii, Honolulu, Hawaii.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

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

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.

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

Liu, S., Yang, J.Y., Zhang, X.Y., Drury, C.F., Reynolds, Hoogenboom, G., 2013. Modelling crop yield, soil water content and soil temperature for asoybean–maize rotation under conventional and conservation tillage systems in Northeast China. Agric. Water Manage. 123, 32–44.

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

Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecol. Modell. 68, 21–32.

McCuen, R.H., Snyder, W.M., 1975. A proposed index for comparing hydrographs. Water Resour. Res. 11, 1021–1024.

Medeiros, P.V., Marcuzzo, F.F.N., Youlton, C., Wendland, E., 2012. Error autocorrelation and linear regression for temperature-based evapotranspiration estimates improvement. J. Am. Water Resour. 48, 297–305.

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

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

Priesack, E., Gayler, S., Hartmann, H.P., 2006. The impact of crop growth sub-model choice on simulated water and nitrogen balances. Nutr. Cycl. Agroecosyst 75, 1–13.

Reckhow, K.H., Clements, J.T., Dodd, R.C., 1990. Statistical evaluation of mechanistic water-quality models. J. Environ. Engine 116, 250–268.

Rinaldi, M., Ventrella, D., Gagliano, C., 2007. Comparison of nitrogen and irrigation strategies in tomato using CROPGRO model. A case study from Southern Italy. Agric. Water Manage. 87 (9), 1–105.

Ritchie, J., Singh. U., Godwin, D., Hunt, L., 1993. A user's guide to CERES Maize-V2.10.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Model 90, 229–244.

Sinclaira, T.R., Seligman, N., 2000. Criteria for publishing papers on crop modeling. Field Crops Res. 68, 165–172.

Smith, P., Smith, J.U., Powlson, D.S., McGill, W.B., Arah, J.R.M., Chertov, O.G., Coleman, K., Franko, U., Frolking, S., Jenkinson, D.S., Jensen, L.S., Kelly, R.H., Klein-Gunnewiek, H., Komarov, A.S., Li, C., Molina, J.A.E., Mueller, T., Parton, W.J., Thornley, J.H.M., Whitmore, A.P., 1997. A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. Geoderma 81, 153–225.

Snedecor, G.W., Cochran, W.G., 1976. Statistical Methods, seventh ed. The Iowa State University Press, Ames, Iowa, USA.

Soler, C.M.T., 2004. Using CERES-Maize model for maize sown off season yield forecast, PhD. Thesis. Escola Superior de Agricultura Luiz de Queiroz, University of Sao Paulo, Piracicaba, Brazil.

Soler, C.M.T., Sentelhas, P.C., Hoogenboom, G., 2005. Thermal time for phenological development of four maize hybrids grown off-season in a subtropical environment. J. Agric. Sci. 143, 169–182.

Soler, C.M.T., Hoogenboom, G., Sentelhas, P.C., Pereira Duarte, A., 2007a. Impact of water stress on maize grown off-season in a subtropical environment. J. Agron. Crop Sci. 193, 247–261.

Soler, C.M.T., Sentelhas, P.C., Hoogenboom, G., 2007b. Application of the CSM-CERES-Maize model for planting date evaluation and yield forecasting for maize grown off-season in a subtropical environment. Eur. J. Agron. 27, 165–177.

Tsuji, G.Y., Uehara, G., Balas, S., 1994. DSSAT v3. University of Hawaii, Honolulu, Hawaii. User's Guide.

Tsuji, G.Y., Hoogenboom, G., Thornton, P.K., 1998. Systems Approaches for Sustainable Agricultural Development vol. 7 Understanding Options for Agricultural Production. Kluwer Academic Publishers in Cooperation with ICASA.

Wagger, M.G., 1983. Nitrogen Cycling in the Plan-Soil System. Ph.D. Thesis, Kansas State University.

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

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

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

Willmott, C.J., Robeson, S.M., Matsuura, K., 2011. Short communication: a refined index of model performance. Int. J. Climatol. 32, 2088–2094.

Yang, J.Y., Huffman, E.C., 2003. EasyGrapher v1.0 Help Manual. In: Jones, J.W., Hoogenboom, G., Wilkens, P.W. et al., Decision Support System for Agrotechnology Transfer Version 4.0, 2003, V3 41-80. University of Hawaii, Honolulu, HI.

Yang, J.Y., Huffman, E.C., 2004. EasyGrapher: software for graphical and statistical validation of DSSAT outputs. Comput. Electron. Agri. 45, 125–132.

Yang, J., Greenwood, D.J., Rowell, D.L., Wadsworth, G.A., Burns, I.G., 2000. Statistical methods for evaluating a crop nitrogen simulation model.NABLE. Agric. Syst. 64, 37–53.

Yang, J.Y., Drury, C.F., Johnston, R., Simard, M., Zavitz, J., Hoogenboom, G., 2010. EasyGrapher v4.5: software for graphical and statistical evaluation of DSSAT v4.5 outputs. Poster presentation. In: Annual meeting of ASA-CSSA-SSSA, 2010, Lang Beach, CA.

Yang, J.M., Dou, S., Yang, J.Y., Hoogenboom, G., Jiang, X., Zhang, Z.Q., Jiang, H.W., Jia, L.H., 2011a. Crop-soil nitrogen cycling and soil organic carbon balance in black soil zone of Jilin province based on DSSAT model. Chin. J. Appl. Ecol. 22, 2075–2083.

Yang, J.M., Liu, J., Dou, S., Yang, J.Y., Hoogenboom, G., 2011b. Evaluation and optimization of best management practices of maize for black soil in Jilin China using the DSSAT model: I. Cultivar calibration and sensitivity analysis of maize yield parameters. Acta Pedologica Sinica. 48, pp. 366–374.