Comparison of statistical indices for the evaluation of crop models performance

ABSTRACT

This study presents a comparison of the usual statistical methods used for crop model assessment. A case study was conducted using a data set from observations of the total dry weight in diploid potato crop, and six simulated data sets derived from the observations aimed to predict the measured data. Statistical indices such as the coefficient of determination, the root mean squared error, the relative root mean squared error, mean error, index of agreement, modified index of agreement, revised index of agreement, modeling efficiency, and revised modeling efficiency were compared. The results showed that the coefficient of determination is not a useful statistical index for model evaluation. The root mean squared error together with the relative root mean squared error offer an excellent notion of how deviated the simulations are in the same unit of the variable and percentage terms, and they leave no doubt when evaluating the quality of the simulations of a model.

Palabras clave:
- Modelo de simulación de cultivos
- Estadísticas de desviación
- Coeficiente de eficiencia
- Índice de concordancia
- Evaluación del modelo
- RMSE

Este artículo presenta una comparación de los métodos estadísticos habituales que se utilizan para la evaluación de modelos de cultivos. Se realizó un estudio de caso utilizando un conjunto de datos observados del peso seco total en un cultivo de papa diploide y seis conjuntos de datos simulados destinados a predecir las observaciones. Los parámetros estadísticos evaluados fueron el coeficiente de determinación, la raíz cuadrada del cuadrado medio del error, la raíz cuadrada del cuadrado medio del error relativo, el error medio, el índice de concordancia, el índice de concordancia modificado, el índice de concordancia revisado, el índice de eficiencia y el índice de eficiencia revisado. Los resultados mostraron que el coeficiente de determinación no es un índice estadístico útil para la evaluación de modelos de cultivo. La raíz cuadrada del cuadrado medio del error junto a la raíz cuadrada del cuadrado medio del error relativo, ofrecen una excelente idea de cuánto están desviadas las simulaciones en la misma unidad de medida de la variable y en términos porcentuales. La raíz cuadrada del cuadrado medio del error y la raíz cuadrada del cuadrado medio del error relativo no dejan dudas al evaluar la calidad de las simulaciones de un modelo respecto a las observaciones.
The traditional research based on field experiments has a high investment in infrastructure, equipment, labor, and time. Alternatives to conventional studies are the development and application of crop models in agriculture, which show a simplified representation of the processes that occur in a real system, including variables that interact and evolve, showing dynamic and real behavior over time (Thornley, 2011). Crop models allow experimentation, complementing traditional research based on field experiments, and allowing an economical and practical evaluation of the effect of different environmental conditions and several agricultural management alternatives, reducing risk, time, and costs (Ewert, 2008).

Several simulation models have been developed for crops such as cassava (Moreno-Cadena et al., 2020), potato (Fleisher et al., 2017; Saqib and Anjum, 2021), wheat (Asseng, 2013; Iqbal et al., 2014), rice (Li et al., 2015), and corn (Abedinpour et al., 2012; Bassu et al., 2014; Kumudini et al., 2014). Moreover, models are continuously evaluated under different environmental conditions, cultivars, and treatments. These crop models are useful tools for simulations of real crop growth and development processes (Yang et al., 2014). The used models are assumptions that have best survived the unremitting criticism and skepticism that are an integral part of the scientific process of construction and development (Thornley, 2011).

In general, the datasets used to develop a crop model are different from the real inputs in which the model is expected to be used. For a crop simulation model to represent a real process, it must be evaluated considering the differences between crop systems, soils, climate, and management practices; otherwise, the conclusions may be speculative and incorrect (Yang et al., 2014).

The growth dynamics represented by crop models are based on a set of hypotheses, which could result in simulation biases or errors (Yang et al., 2014). Thus, the model performance evaluation is crucial by comparing model estimates to actual values, and this process includes a criteria definition that relies on mathematical measurements of how well the estimates produced by the model simulate the observed values (Ramos et al., 2018). This statistical analysis is considered as the critical method to compare the model outputs with the measured data (Montoya et al., 2016; Reckhow et al., 1990; Willmott et al., 1985; Yang et al., 2000).

The most common methods for assessing the reliability of simulation models are based on the analysis of differences between measured and simulated values, and on regression analysis, also between measured and simulated values (Lin et al., 2014; Willmott, 1982; Yang et al., 2000). However, many authors who research crop modeling use such methods without detailing methodological basis and using terminology and symbols that create confusion. For example, in the analysis of the difference, statistics such as relative error (RE), index of agreement (d), and modeling efficiency (EF) may be useful when comparing the simulation capability of one model with another, but not when comparing what is observed with what is simulated in the same model (Ramos et al., 2018; Yang et al., 2014). Relative error (RE), which relates the error between measured and simulated values, concerning the measured average, represents the relative size of the average difference (Willmott, 1982), indicating whether the magnitude of the root-mean-square error (RMSE) is low, medium or high. However, it has the disadvantage that it can be affected by the magnitude of the values, by outliers, and the number of observations. It may be the case that two groups of data with high and low values, present a similar RMSE. However, having different averages, RE values will also be different (Cao et al., 2012).

Because of its simplicity, regression analysis is often misused to evaluate simulation models. In some cases, the RMSE that measures the average difference between measured and simulated values tends to be used indiscriminately, without considering that it is different from the RMSE obtained in regression analysis (Willmott, 1982). The coefficient of determination ($R^2$) is a measure of the linear regression adjustment, which, when used in isolation, makes no sense since the goal is to evaluate the crop simulation model, not the regression model obtained.

The magnitude of $R^2$ does not necessarily reflect whether the simulated data represent well the observed data since it is not consistently related to the accuracy
Comparison of statistical indices for the evaluation of crop models performance

of the prediction (Willmott, 1982). This is because an $R^2$ can be obtained close to 1.0 but below or above the 1:1 line, tending to simulate high values or underestimates the observed values, respectively.

Many statistical indices are frequently used in model evaluation, and this paper aimed to compare and improve the understanding and interpretation of these conventional statistical indices in a case study.

MATERIALS AND METHODS

The performance of nine statistical indices was computed to evaluate the simulations of actual observations and simulations of total dry weight (kg ha$^{-1}$) obtained in a diploid potato field experiment conducted in Medellín, Colombia. This data set were taken from Saldaña-Villota and Cotes-Torres (2020). Besides, from the actual observed data, six data sets were generated with arbitrary deviations appropriately imposed to illustrate the behavior of the statistical indices under evaluation. (Table 1). In case 1, the first half of the simulations is overestimated, and the second half is underestimated in the same amount (200 kg ha$^{-1}$). In case 2, the first half of the simulations is overestimated 1.5 times, and the second half of the simulations is underestimated 0.5 times. In case 3, all simulations are overestimated in 100 kg ha$^{-1}$. In case 4, all simulations are overestimated 2.5 times. In case 5, most of the simulations are overestimated in different proportions, and an outlier 3.4 times larger than its corresponding observation is presented. Finally, in case 6, all simulations are overestimated in different proportions, and they do not have any relationship with the observations.

The indices are expected to inform the researcher of the accuracy of any model in simulating the observations.

Table 1. Actual observations of diploid potato total dry weight (kg ha$^{-1}$) and simulated data sets.

<table>
<thead>
<tr>
<th>Days after planting (DAP)</th>
<th>Actual observed data set</th>
<th>Simulated Data Sets Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>57.43</td>
<td>257.44</td>
</tr>
<tr>
<td>30</td>
<td>153.20</td>
<td>353.20</td>
</tr>
<tr>
<td>37</td>
<td>315.10</td>
<td>515.10</td>
</tr>
<tr>
<td>43</td>
<td>547.59</td>
<td>747.59</td>
</tr>
<tr>
<td>51</td>
<td>804.70</td>
<td>1004.70</td>
</tr>
<tr>
<td>58</td>
<td>1166.00</td>
<td>1366.00</td>
</tr>
<tr>
<td>65</td>
<td>1338.00</td>
<td>1138.00</td>
</tr>
<tr>
<td>72</td>
<td>1837.00</td>
<td>1637.00</td>
</tr>
<tr>
<td>79</td>
<td>2740.00</td>
<td>2540.00</td>
</tr>
<tr>
<td>85</td>
<td>4103.00</td>
<td>3903.00</td>
</tr>
<tr>
<td>91</td>
<td>5657.00</td>
<td>5457.00</td>
</tr>
<tr>
<td>100</td>
<td>6738.00</td>
<td>6538.00</td>
</tr>
<tr>
<td>Mean</td>
<td>2121.42</td>
<td>2121.42</td>
</tr>
</tbody>
</table>

$^a$ Total dry weight in diploid potato crop. Source: Saldaña-Villota and Cotes-Torres (2020).

Case 1: The first half of the simulations is overestimated, and the second half is underestimated in the same amount (200 kg ha$^{-1}$).
Case 2: The first half of the simulations is overestimated 1.5 times, and the second half of the simulations is underestimated 0.5 times.
Case 3: All simulations are overestimated in 100 kg ha$^{-1}$.
Case 4: All simulations are overestimated 2.5 times.
Case 5: Most of the simulations are overestimated in different proportions, and an outlier 3.4 times larger than its corresponding observation is presented.
Case 6: All simulations are overestimated in different proportions, and they do not have any relationship with the observations.
The statistical indices are expected to allow decisions to be made regarding the acceptance or rejection of the models. In this study, with the modifications applied to generate the six cases, the statistical indices must accept cases 1 and 3 and reject the other cases without ambiguity.

Many statistical indices are commonly used in model evaluation, and they have been classified depending on their mathematical formulation. In this study, nine indexes were evaluated and classified into two categories. The first one corresponds to the 'test statistics', and the second one corresponds to measures of accuracy and precision called 'deviation statistics' (Ali and Abustan, 2014; Willmott et al., 1985; Yang et al., 2014).

### Test statistics

Linear regression and coefficient of determination ($R^2$) are used to explain how well the simulations ($y$) represent the observations ($x$) (Kobayashi and Salam, 2000; Moriasi et al., 2007; Willmott, 1982). The linear model follows Equation 1.

$$y = \alpha + \beta x + \varepsilon$$  \hspace{1cm} (1)

where $\alpha$ is the regression intercept, $\beta$ is the slope, and $\varepsilon$ represents the random error.

The $R^2$ assesses the goodness of fit of the linear model by measuring the proportion of variation in $y$, which is accounted for by the linear model. $R^2=1.0$ indicates a perfect fit of Equation 1, and $R^2=0$ means there is no linear relationship.

However, many researchers have reported the limitation of $R^2$ in the appropriate evaluation of the models, remarking that $R^2$ estimates the linear relationship between two variables, and it is no sensitive to additive and proportional differences between model estimates and measured data (Kobayashi and Salam, 2000; McCuen and Snyder, 1975; Willmott, 1981). The authors also indicate that the relationship may be non-linear, which would be an additional problem.

### Deviation statistics

Some deviation statistics correspond to measures developed to test the deviation directly (deviation = $y-x$) and surpass the limitation of correlation-based statistics (Yang et al., 2014). The Mean Error ($E$) (Equation 2) indicates whether the model simulations ($y$) overestimate or underestimate the observations ($x$). When $E>0$ means that the model is overestimating, while $E<0$ means that model underestimates the measured data. $E$ has a drawback: the positive and negative errors can negate each other, and large positive and negative deviations can still obtain $E=0$ (Addiscott and Whitmore, 1987; Yang et al., 2000).

$$E = n^{-1} \sum_i (y_i - x_i)$$  \hspace{1cm} (2)

where $i=1,2,...,n$.

Due to $E$ disadvantage, some measures based on the sum of squares were developed (Yang et al., 2014). The root mean square error ($RMSE$) (Equation 3) has the same unit of deviation $y-x$, and it is frequently used in both model calibration and validation process (Hoogenboom et al., 2019; Hunt and Parsons, 2011)

$$RMSE = \left[ \frac{1}{n} \sum_i (y_i - x_i)^2 \right]^{0.5}$$  \hspace{1cm} (3)

The relative root mean square error ($rRMSE$) (Equation 4) is a relative measure used for comparisons of different variables or models. indicating whether the magnitude of the root-mean-square error ($RMSE$) is low, medium, or high (Priesack et al., 2006).

$$rRMSE = \frac{RMSE}{\bar{x}} \times 100$$  \hspace{1cm} (4)

Nash-Sutcliffe modeling efficiency coefficient ($EF$) (Equation 5) (Nash and Sutcliffe, 1970). This index is a dimensionless measure ($-\infty$ to 1.0). A perfect fit between simulations and observations produces an $EF=1.0$. Any value between 0 and 1.0 is obtained for any realistic simulation. $EF<0$ is obtained if the simulated values are worse than merely using the observed mean $\bar{x}$ to replace the simulated $y_i$.

$$EF = 1 - \frac{\sum_i (y_i - x_i)^2}{\sum_i (x_i - \bar{x})^2}$$  \hspace{1cm} (5)

Another index that is commonly used in crop model evaluation is the index of agreement ($d$) (Equation 6) a dimensionless measure (0 to 1.0) proposed by Willmott (1982). This index has been recommended by researchers in modeling to carry out comparisons between simulated values and measured data (Krause et al., 2005; Moriasi et al., 2007).
Comparison of statistical indices for the evaluation of crop models performance

**RESULTS AND DISCUSSION**

This study shows a comparison of nine statistical indexes used during model evaluation. The actual data of the total dry weight measured in a diploid potato field experiment and the six simulated data set are shown in Figure 1 to facilitate the visualization of the data and their analyzes.

**Coefficient of determination ($R^2$)**

In the simulated data cases 1, 2, 3, and 4 (Figure 1A-D), different scenarios were presented in which the actual observations are overestimated or underestimated. The simulations preserved the trend of the measurements, which is the reason why the $R^2$ was high. Although simulations considerably overestimated the measurements in case 4, the fact that the simulated data follow the trend of the observations even if they are overestimated or underestimated, the $R^2$ will be close to 1.0. Consequently, this index is not adequate to evaluate the quality of the simulations in growth variables in crop models. The coefficient of determination was lower in cases 5, and 6 (Figures 1E and F), indicating that the simulated data did not follow the observed data trend.

**Mean error ($E$)**

$E$ indicates whether the model overestimates or underestimates the measurements. This index presented difficulty to indicate what happened in case 1, in which half of the simulations were overestimated, and half were underestimated in the same proportion. In this case, $E=0$, and this value gives no indication of over or underestimation. In case 2, $E$ indicates that the simulated data underestimate the total dry weight by 807,040 kg ha$^{-1}$. In the remaining cases, $E>0$, indicating that the simulations overestimate the measurements. According to $E$, case 6 was the one that registered the maximum overestimation, exceeding 6000 kg ha$^{-1}$.

**Root mean squared error ($RMSE$) and the relative-RMSE ($rRMSE$)**

The $RMSE$ indicates how deviated the simulated mean is from the observed mean. This index does not indicate whether there are overestimates or underestimates. Nevertheless, if the $RMSE$ is close to zero or less than the amount assigned by the researcher according to the expertise in the crop studied, the model performs better in predicting the measured data. If the researcher is not an

\[ d = 1 - \frac{\sum (y_i - x_i)^2}{\sum (|y_i - \bar{x}| + |x_i - \bar{x}|)^2} \]  

\[ EF = 1 - \frac{\sum |y_i - x_i|}{\sum |x_i - \bar{x}|} \]  

\[ d = 1 - \frac{\sum (y_i - x_i)^2}{\sum (|y_i - \bar{x}| + |x_i - \bar{x}|)^2} \]  

\[ d' = 1 - \frac{\sum |y_i - x_i|}{2 \sum |x_i - \bar{x}|} \]  

The calculation of the statistics indices to evaluate the six simulated data sets, and figures were made with R statistical software (R Core Team, 2020).
Figure 1. Comparison between real observations and six simulated data set of total dry weight in diploid potato crop (kg ha$^{-1}$) over time (days after planting). Black circles correspond to the real observations, and red ones correspond to the simulated counterpart.
Comparison of statistical indices for the evaluation of crop models performance

According to Nash and Sutcliffe (1970), $E_F$ and $r$-$R$-$M$-$S$ values between 0 and 1 are expected in any modeling scenario. However, in case 2, for instance, $E_F$ and $r$-$R$-$M$-$S$ reached values of 0.506 and 0.408, which are values higher than zero, but by themselves, they are not clear with the reality of the simulation quality. Nonetheless, values less than zero in these two indices are indicators of wrong predictions; thus, cases 2, 5, and 6 achieved values <0, confirming what $E$, $R$-$M$-$S$, and $r$-$R$-$M$-$S$ had indicated. Also, the more negative values suggest that the simulated data were worse. The clearest example is case 6, which reached $-15.689$ in $E_F$, but $E_F$ was $-2.531$. $E_F$ reached higher values (both positive and negative) because when considering sums in terms of the sum of squares in its formulation, it is more affected by outliers. $E_F$ is calculated considering the sum in terms of absolute values, that means less sensitivity to extreme data.

**Index of agreement ($d$), modified index of agreement ($d'$), and revised index of agreement ($d''$)**

Finally, from the group of indices $d$, $d'$, and $d''$, the best simulations reach values close to 1.0 (Cases 1 and 3). In the same way as $E_F$ and $r$-$R$-$M$-$S$, the statistics of group $d$ must be estimated in association with other indices to make better inferences about the accuracy of the simulation. In case 2, $d$=0.764, and if this value is analyzed by itself, it would suggest that the model is adequate, but $d'$ is stricter than $d$, and its value is clearer suggesting that the simulations are not adequate ($d''$=0.637). $d$ and $d'$ in cases 5 and 6 were less than 0.75, suggesting that these models are not suitable for simulating the measured data set. Case 6 was the only one that reached a negative $d''$ value ($-0.765$), again indicating that the simulations, in this case, are not adequate when predicting the measurements.

**General performance of the statistical indices in evaluating the quality of the simulations of a model**

Summarizing the previous results (Table 2), the $R$-$M$-$S$, $r$-$R$-$M$-$S$, $E_F$, $d$, and $d''$ are the best indices for evaluating the quality of simulations because,
they accepted the simulations in cases 1 and 3, and rejected the other cases, which was expected in this study when comparing the behavior of the statistical indices. However, given the simplicity in the interpretation of RMSE and rRMSE, they are preferred over dimensionless statistics.

Table 2. Acceptance or rejection of the simulations defined by different statistical indices for each data set.

<table>
<thead>
<tr>
<th>Statistical parameter</th>
<th>Simulated Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$R^2$</td>
<td>✓</td>
</tr>
<tr>
<td>$E$</td>
<td>-</td>
</tr>
<tr>
<td>RMSE</td>
<td>✓</td>
</tr>
<tr>
<td>rRMSE</td>
<td>✓</td>
</tr>
<tr>
<td>$EF$</td>
<td>✓</td>
</tr>
<tr>
<td>$EF_1$</td>
<td>✓</td>
</tr>
<tr>
<td>$d$</td>
<td>✓</td>
</tr>
<tr>
<td>$d_1$</td>
<td>✓</td>
</tr>
<tr>
<td>$d_1'$</td>
<td>✓</td>
</tr>
</tbody>
</table>

✓ Simulations accepted  
X Simulations rejected

Statistical analysis is a crucial procedure during model calibration and evaluation, and there are many statistical methods useful to support crop model researchers. It is unquestionably that $R^2$ is not a suitable parameter for model evaluation because it is not sensitive to additive (regression intercept) and proportional differences (regression slope) (Willmott et al., 2012; Yang et al., 2013). The linear regression should be employed to evaluate the simulated outputs with the observed inputs when the time series data follow the assumptions of independence, normality, and homoscedasticity in the error term (Yang et al., 2014). The error term does not follow these assumptions in the deviation statistics because they are not hypothesis tests (Willmott et al., 1985).

The mean error ($E$) is a good statistical parameter to quickly determine if the model under or overestimates the observations. Unfortunately, it does not offer clarity on the quality of the simulations. Nevertheless, RMSE and rRMSE are very suitable for model evaluation because they provide the researcher with a useful decision-making guide. It is important to highlight the advantages that the RMSE and the rRMSE offer, which together offer a better idea of how deviated the simulations are in the same unit of the variable and percentage terms.

If only the group index of agreement is considered during the evaluation of a model, it is possible to make bad decisions or assume that the model predicts the measured data with quality when in reality, the predictions are not adequate. $d$ can quickly reach 1.0 without considering significant discrepancies between simulations and observations because the sum of squares-based deviations easily inflates $d$. A researcher could consider case 2 a suitable model to simulate the observations according to $d$ and $d_1'$ values, even when $d_1'$ seems to be stricter than $d$ in mathematical terms. $d_1$ and $EF$ showed well behaviors, and they have a sharp meaning and interpretation when values tend to zero. Yang et al. (2014) suggested for plant growth variables simulations $EF>0$ and $d$, $d_1$, and $d_1'$ as minimum values for dry weight of leaves, stems, yield, tubers, total in the case of the potato crop.

Both modeling efficiency coefficients ($EF$ and $EF_1$) and indices of agreement ($d$, $d_1$, and $d_1'$) are widely used in modeling evaluation. Although $d$ and $EF$ are sensitive to the sum of squares and, in consequence, they achieve higher values even with not accurate simulations. The researcher should use these dimensionless indices carefully. Alternatively, use RMSE and rRMSE as good guides to evaluate the quality of the models.

CONCLUSION
The RMSE and the rRMSE offer a better idea of how deviated the simulations are in the same unit of the variable and percentage terms; for this reason, these
indices are the most appropriate to reflect the quality of the simulations of a model. This pair of indices was the only one that unquestionably established that cases 1 and 3 are almost perfect with deviations less than 200 kg ha⁻¹, which is less than 10% concerning the mean of the observations. RMSE and rRMSE leave no doubt that cases 2, 4, 5, and 6 correspond to models that reflect very poorly or do not reflect the observations.

REFERENCES


