DOI: http://dx.doi.org/10.1590/1678-4499.2017054

AGROMETEOROLOGY - Article

On the performance of three indices of agreement: an easy-to-use r-code for calculating the Willmott indices

Heloisa Ramos Pereira, Monica Cristina Meschiatti, Regina Célia de Matos Pires, Gabriel Constantino Blain*

Instituto Agronômico - Centro de Ecofisiologia e Biofísica - Campinas (SP), Brazil.

ABSTRACT: A key step for any modeling study is to compare model-produced estimates with observed/reliable data. The original index of agreement (also known as original Willmott index) has been widely used to measure how well model-produced estimates simulate observed data. However, in its original version such index may lead the user to erroneously select a predicting model. Therefore, this study compared the sensibility of the original index of agreement with its two newer versions (modified and refined) and provided an easy-to-use R-code capable of calculating these three indices. First, the sensibility of the indices was evaluated through Monte Carlo Experiments. These controlled simulations considered different sorts of errors (systematic, random and systematic + random) and errors magnitude. By using the R-code, we also carried out a case of study in which the indices are expected to indicate that the empirical Thornthwaite's model produces poor estimates of daily reference evapotranspiration in respect to the standard method Penman-Monteith (FAO56). Our findings indicate that the original index of agreement may indeed erroneously select a predicting model performing poorly. Our results also indicate that the newer versions of this index overcome such problem, producing more rigorous evaluations. Although the refined Willmott index presents the broadest range of possible values, it does not inform the user if a predicting model overestimate or underestimate the simulated data, resulting in no extra information regarding those already provided by the modified version. None of the indices represents the error as linear functions of its magnitude in respect to the observed process.

Key words: modified index of agreement, refined index of agreement, model performance.

*Corresponding author: gabriel@iac.sp.gov.br

Received: Feb. 15, 2017 – Accepted: May 29, 2017

INTRODUCTION

The great capacity of modern personal computers has enabled the use of complex models to simulate and describe increasing numbers of natural and man-made processes. In this view, numerical models of environmental, hydrological and agro-meteorological systems have grown in number and complexity (Willmott et al. 2012). Accordingly, the evaluation of model performance, i.e., to compare model-produced estimates with observed/reliable values, is a fundamental step for model development and use (Willmott et al. 1985; Willmott et al. 2012). This validation process commonly includes a criteria definition that relies on mathematical measurements of how well model-produced estimates simulate the observed values (Willmott et al. 1985; Krause et al. 2005; Willmott et al. 2012).

Willmott (1981; 1982) and Willmott and Wicks (1980) proposed and used an index of agreement – currently referred to as 'the Willmott index', 'the original d index' or simply 'the d index' (d_{orig}) – that is intended to be a dimensionless measurement of model accuracy. The d_{orig} is bounded by 0, meaning no agreement, and 1, meaning a perfect fit (Willmott 1984). Authors such as Legates and McCabe (1999) stated that d_{orig} represented a remarkable improvement in respect to the coefficient of determination. In this view, the d_{orig} index has been used in several meteorological, agrometeorological and hydrological studies (e.g. Wu et al. 2005; Meschiatti and Blain 2016). In spite of this widespread use, Willmott et al. (1985) noted that the use of squared differences in its calculation algorithm might result in high values of this index ($d_{orig} \approx 1$) even in the presence of large errors. In addition, sums-of-squares-based measurements vary in response to both variability and central tendency within a set of deviations (Willmott et al. 2009). On such background, Willmott (1984) proposed a modification in the d_{orig} index that replaces the square function by the modulus of the deviations. This modified version is frequently referred to as the modified index of agreement (d_{mod}). The advantages of d_{mod} over d_{orig} is that errors and differences are given their appropriate weighting factors (e.g. Willmott et al. 1985; Willmott et al. 2009). The d_{mod} may be regarded as a more rigorous method than d_{orig} (Bardin-Camparotto et al. 2013) because when these two indices are applied to the same validation process, d_{mod} tends to approach its maximum value more slowly as the predicted values approach the observed data (Legates and McCabe 1999; Willmott et al. 2012).

In spite of the advantages of d_{mod} over its original version, Willmott et al. (2012) stated that the overall range of d_{orig} and d_{mod} [0:1] is narrow to adequately represent the great variety of forms that predicted values can differ from observed data. Therefore, these authors proposed a new index, referred to as the refined index of agreement (d_{ref}), that is bounded by –1.0 and 1.0. Willmott et al. (2012) claimed that the d_{ref} is more rationally related to model accuracy than d_{orig} and d_{mod} (Willmott et al. 2012).

Despite the above-mentioned efforts to improve the original version of the d_{orig} index, an overview on the scientific literature indicates that the use and evaluation of d_{mod} and d_{ref} still need to be enhanced. This statement is particularly true for tropical developing countries that use d_{orig} as a tool for crop modeling and other agrometeorological studies. Therefore, in order to motive the use of the two modified versions of the index of agreement, the goals of this study were (i) to evaluate and compare the performance of d_{orig}, d_{mod} and d_{ref} to different sorts of errors (systematic, random and systematic+random) and (ii) to provide an easy-to-use R-code capable of calculating the three indices.

MATERIAL AND METHODS

The performance of the three indices was evaluated (i) by means of controlled Monte Carlo experiments, and (ii) by means of a well-documented case of study in which the empirical Thornthwaite's model is used to estimate daily amounts of reference evapotranspiration (ETo). As it will be further described, the indices are expected to inform the user that this empirical model tends to produce poor estimates of daily ETo values. The Willmott indices (d_{orig}, d_{mod} and d_{ref}) can be calculated as follow:

$$d_{orig} = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \quad (1)$$

$$d_{mod} = 1 - \frac{\sum_{i=1}^{n}|P_i - O_i|}{\sum_{i=1}^{n}(|P_i - \bar{O}| + |O_i - \bar{O}|)} \quad (2)$$

$$d_{ref} = \begin{cases} 1 - \frac{\sum_{i=1}^{n}|P_i - O_i|}{2\sum_{i=1}^{n}|O_i - \bar{O}|}, \text{when} \\ \sum_{i=1}^{n}|P_i - O_i| \leq c \sum_{i=1}^{n}|O_i - \bar{O}| \\ \frac{2\sum_{i=1}^{n}|O_i - \bar{O}|}{\sum_{i=1}^{n}|P_i - O_i|} - 1, when \\ \sum_{i=1}^{n}|P_i - O_i| > 2 \sum_{i=1}^{n}|O_i - \bar{O}| \end{cases} \quad (3)$$

where *P* and *O* are, respectively, the predicted and observed values and the 95% confidence interval of each index value were estimated through the bootstrap approach, as recommended by Willmott et al. (1985).

Monte Carlo Simulation: Systematic errors

All sets of Monte Carlo simulations were performed using a sample size N = 100. This N value was adopted to avoid the influence of different sample sizes on the outcomes of the simulations. Naturally, the effect of different N values on d_{orig}, d_{mod} and d_{ref} should be addressed in future studies. The first set of simulations evaluated the performance of the three indices in the presence of systematic errors. In order to cover a great range of behaviors commonly found in agrometeorological variables (Blain 2014), the observed values (o_i) were generated using the gamma 2-parameter distribution (expression 4). As highlighted by authors such as Wilks (2011), the gamma distribution can assume several shapes, depending on the value of its parameters. This flexible distribution either can assume the exponential form or can approach the bell-shaped form of the Gaussian distribution (Wilks 2011). Therefore, the shape (α) and scale (β) parameters of this distribution were respectively set to the following values: G(1,30); G(1,50); G(2,30); G(2,50); G(4,30); G(4,50). Within the R-software environment, gamma-distributed variables can be generated as follows.

```
o = as.vector(rgamma(N, α,1/β))                    (4)
```

The predicted data (p_i) were generated from the observed values adding a systematic error proportional to the mean value of the o_i series (mean (o_i); expression 5) so that the magnitude of the systematic deviations, in respect to the mean o_i values, were 10%, 20% and 30%.

```
p = as.vector(o+(mean(o)*CF))                      (5.1)

p = as.vector(o-(mean(o)*CF))                      (5.2)
```

where CF is the change factor and it was set to the values of 0.1, 0.2 and 0.3.

Monte Carlo Simulation: Random errors

In this section, the predicted data were generated from the same observed values generated in section "Systematic errors". However, instead of systematic deviations, we added normally distributed random errors proportional to the process mean value, as described in expression 6. The change factor assumed the same values as those of the section "Systematic errors" (p and o have the same mean value).

```
p = as.vector(o+(rnorm(N,0,1)*(mean(o)*CF)))       (6.1)

p = as.vector(o-(rnorm(N,0,1)*(mean(o)*CF)))       (6.2)
```

Monte Carlo Simulation: Random and Systematic errors

In this section, the predicted values were generated from the same observed values described in section Systematic errors" . Both systematic and random deviations were added according to expression 7. The change factor assumed the same magnitudes as those of the two previous sections.

```
p = as.vector(o+(rnorm(N,0,1)*(mean(o)*            (7.1)
   (change/2)) + (mean(o)*(change/2))))

p = as.vector(o-(rnorm(N,0,1)*(mean(o)*            (7.2)
   (change/2))-(mean(o)*(change/2))))
```

All Monte Carlo simulations were performed considering errors equal to or lower than 30% of the process mean value. This limit was adopted based on the assumption that models leading to errors equal to or larger than this threshold would be dismissed without the need to use a measurement of agreement. A simple visual analysis of the estimated values would indicate that the prediction model is performing poorly.

Easy-to-use R-code

In order to motive the use of the Willmott's indices, we developed a computational algorithm by using the R-software, which is a free environment for graphics and statistical computing (www.r-project.org). The code was developed so that practically no previous knowledge about the software is required. Naturally, advanced users can easily modify the code according to their needs. The code is described in Table 1.

As a case of study, this code was applied to compare the performance of the empirical Thornthwaite's model (TW) in estimating daily amounts of reference evapotranspiration (ETo) in Campinas, State of São Paulo, in respect to those estimated from physically-based Penman-Monteith method (Allen et al. 1998). The Climatic variability in this location →

Table 1. R-code for calculating the Willmott's indices.

```
setwd("C:/Mydatafile") #set a working directory called Mydatafile.
#data.txt is a 2 column data file.
#The 1st column is the observed while the 2nd column is the predicted values.
data=as.matrix(read.table("data.txt", head=T)) # If the data file has no head head=F
o=as.matrix(data[,1])
p=as.matrix(data[,2])
N=length(o)
databoot=matrix(NA,N,2)
Nboots=10000
dorigboot=matrix(NA,Nboots,1)
drefboot=matrix(NA,Nboots,1)
dmodboot=matrix(NA,Nboots,1)
MAEboot=matrix(NA,Nboots,1)
alfa=0.05 #significance level: e.g. 0.05 for 5%; 0.1 for 10%
# MAE
MAE=sum((abs(p-o)))/N
Num=sum(abs(o-p)); Numorig=sum((o-p)^2)
Den=sum(abs(p-mean(o))+abs(o-mean(o)))
Denorig=sum((abs(p-mean(o))+abs(o-mean(o)))^2)
# Original Willmott index
dorig=1-((Num^2)/Den^2)
# Modified Willmott index
dmod=1-(Num/Den)
# Refined Willmott's index
if (abs(sum(p-o))<=2*sum(abs(o-mean(o)))){
dref=1-((sum(abs(p-o)))/(2*sum(abs(o-mean(o))))} else
{(dref=((2*sum(abs(o-mean(o))))/sum(abs(p-o)))-1)}
# Bootstraping
alfa1=alfa/2
for (i in 1:Nboots){
databoot=data[sample(nrow(data),replace=TRUE),]
oboot=as.matrix(databoot[,1])
pboot=as.matrix(databoot[,2])
# MAE
MAEboot[i,1]=sum((abs(pboot-oboot)))/N
Numboot=sum(abs(oboot-pboot)); Numorigboot=sum((oboot-pboot)^2)
Denboot=sum(abs(pboot-mean(oboot))+abs(oboot-mean(oboot)));
Denorigboot=sum((abs(pboot-mean(oboot))+abs(oboot-mean(oboot)))^2)
# Original Willmott index
dorigboot[i,1]=1-((Numboot^2)/Denboot^2)
# Modified Willmott index
dmodboot[i,1]=1-(Numboot/Denboot)
# Refined Willmott's index
if (abs(sum(pboot-oboot))<=2*sum(abs(oboot-mean(oboot)))){
drefboot[i,1]=1-((sum(abs(pboot-oboot)))/(2*sum(abs(oboot-mean(oboot))))} else
{(drefboot[i,1]=((2*sum(abs(oboot-mean(oboot))))/sum(abs(pboot-oboot)))-1)}}
#Defining confidence intervals
MAE_Clinf=quantile(MAEboot, probs=alfa1, na.rm=T)
MAE_Clsup=quantile(MAEboot, probs=(1-alfa1), na.rm=T)
dorig_Clinf=quantile(dorigboot, probs=alfa1, na.rm=T)
dorig_Clsup=quantile(dorigboot, probs=(1-alfa1), na.rm=T)
dmod_Clinf=quantile(dmodboot, probs=alfa1, na.rm=T)
dmod_Clsup=quantile(dmodboot, probs=(1-alfa1), na.rm=T)
dref_Clinf=quantile(drefboot, probs=alfa1, na.rm=T)
dref_Clsup=quantile(drefboot, probs=(1-alfa1), na.rm=T)
ModelAcuracy=cbind(MAE_Clinf,MAE,MAE_Clsup,dorig_Clinf,dorig,dorig_Clsup,d
mod_Clinf,dmod,dmod_Clsup,dref_Clinf,dref,dref_Clsup)
print(ModelAcuracy)
```

is influenced by monsoon system (Carvalho et al. 2004), in which the wet season occurs during the austral summer associated with the South Atlantic Convergence Zone. In the winter, the high pressing system of the South Atlantic leads to climatically dry conditions. Regarding the performance of the two above-mentioned models, it is well documented that the TW model is not suited for estimating ETo values at daily scale (Carvalho et al. 2011). On the other hand, due to its solid physical fundamentals, the PM model (Allen et al. 1998) is regarded as the standard method for estimating EP amounts at daily scale. Therefore, the Willmott's indices should indicate significant differences between ETo values estimated from these two methods. The Willmott indices were applied to daily ETo values (December 2007 to November 2009; Campinas-SP) within each season – Summer (December to February), Fall (March to May), Winter (June to August) and Spring (September to November).

RESULTS AND DISCUSSION

The analysis of the Monte Carlo simulations must first take into account an important difference between d_{orig} and the other two versions of the Willmott's indices. As can be noted from Equations 1, 2 and 3, only the original index (Equation 1) squares the errors ($o_i - p_i$). However, when applied to large errors magnitude, the square function increases the influence of these deviations on the sum-of-squared errors (Willmott 1982; 1984; Willmott et al. 1985; Legates and McCabe 1999; Willmott et al. 2012). In practical terms, the result of such feature is that high d_{orig} values may be obtained in the presence of relatively large errors. Considering all Monte Carlo Simulations performed in this study, no d_{orig} value lower than 0.80 was observed. This statement holds for all sorts of errors evaluated in this study (Figures 1 to 4), indicating that relatively high d_{orig} values may be observed in the presence of a prediction model performing poorly (Willmott et al. 1985; Legates and McCabe 1999; Willmott et al. 2012; Bardin-Camparotto et al. 2013). Therefore, the findings of this study support the statement that both d_{ref} and d_{mod} are more rigorous than d_{orig}, and that they should be preferred over their original version.

The results of the Monte Carlo simulation also indicated that errors with the same magnitude but opposite signs (expressions 4.1 against 4.2; 5.1 against 5.2; 6.1 against 6.2) lead both d_{mod} and d_{orig} to assume a unique value. This feature

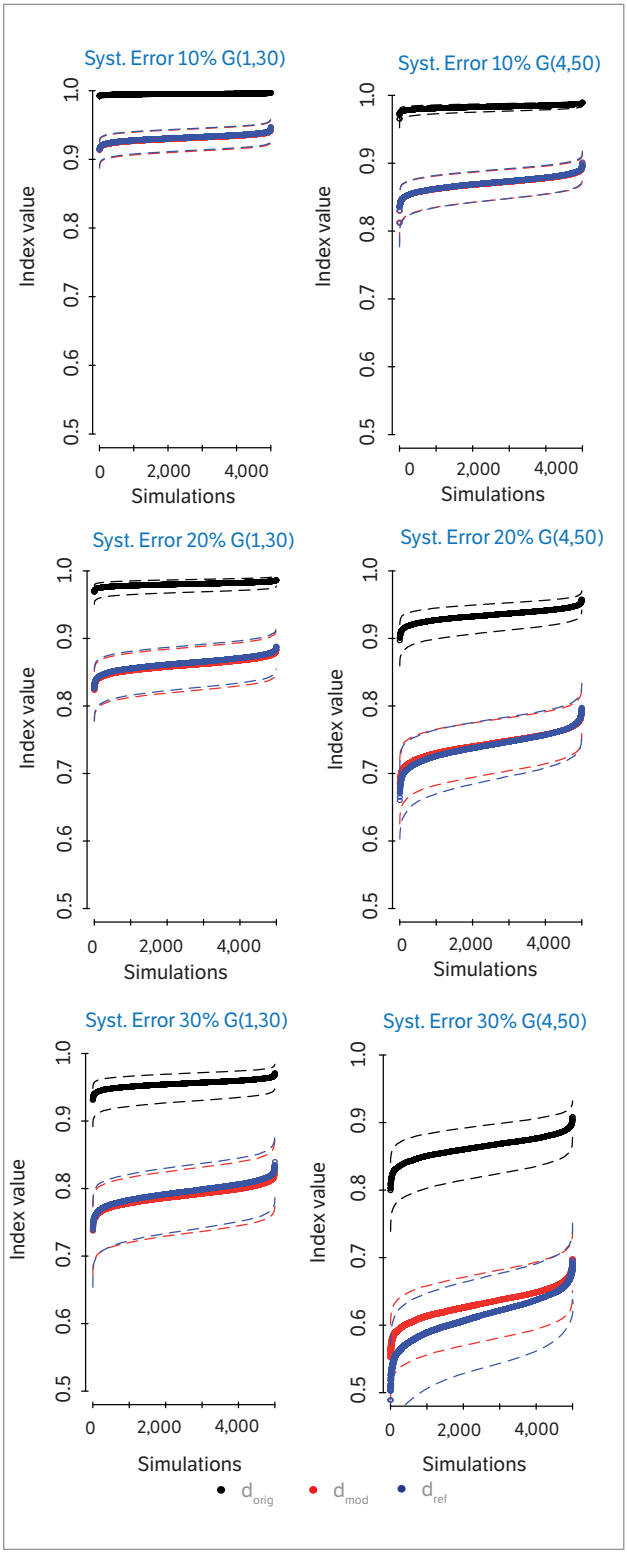

Figure 1. Performance of the original (d_{orig} black dots), modified (d_{mod} red dots) and refined (d_{ref} blue dots) indices of Agreement subjected to (positive) systematic errors. The observed values were generated from a 2-parameter gamma distribution [G(.)] with shape parameter set to 1 and 4 and scale parameter set to 30 and 50.

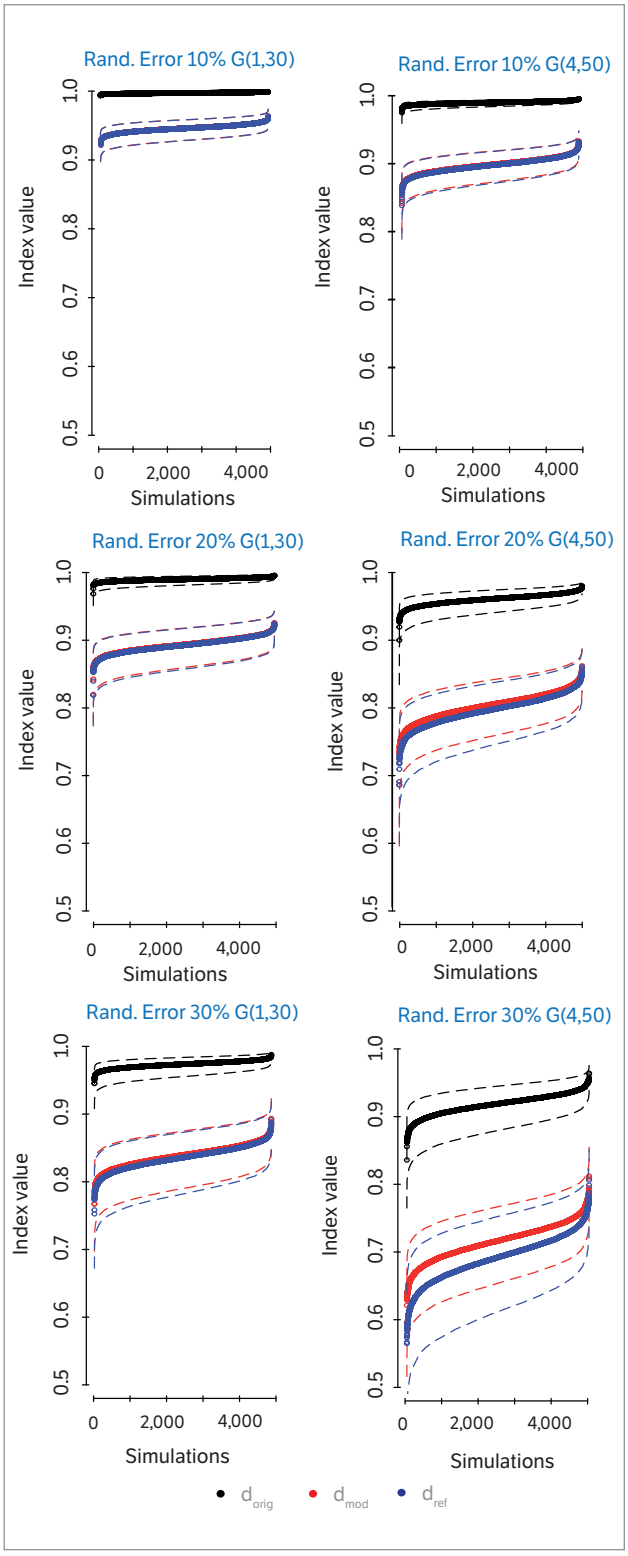

Figure 2. Performance of the original (d_{orig} black dots), modified (d_{mod} red dots) and refined (d_{ref} blue dots) indices of Agreement subjected to (positive) random errors. The observed values were generated from a 2-parameter gamma distribution [G(.)] with shape parameter set to 1 and 4 and scale parameter set to 30 and 50.

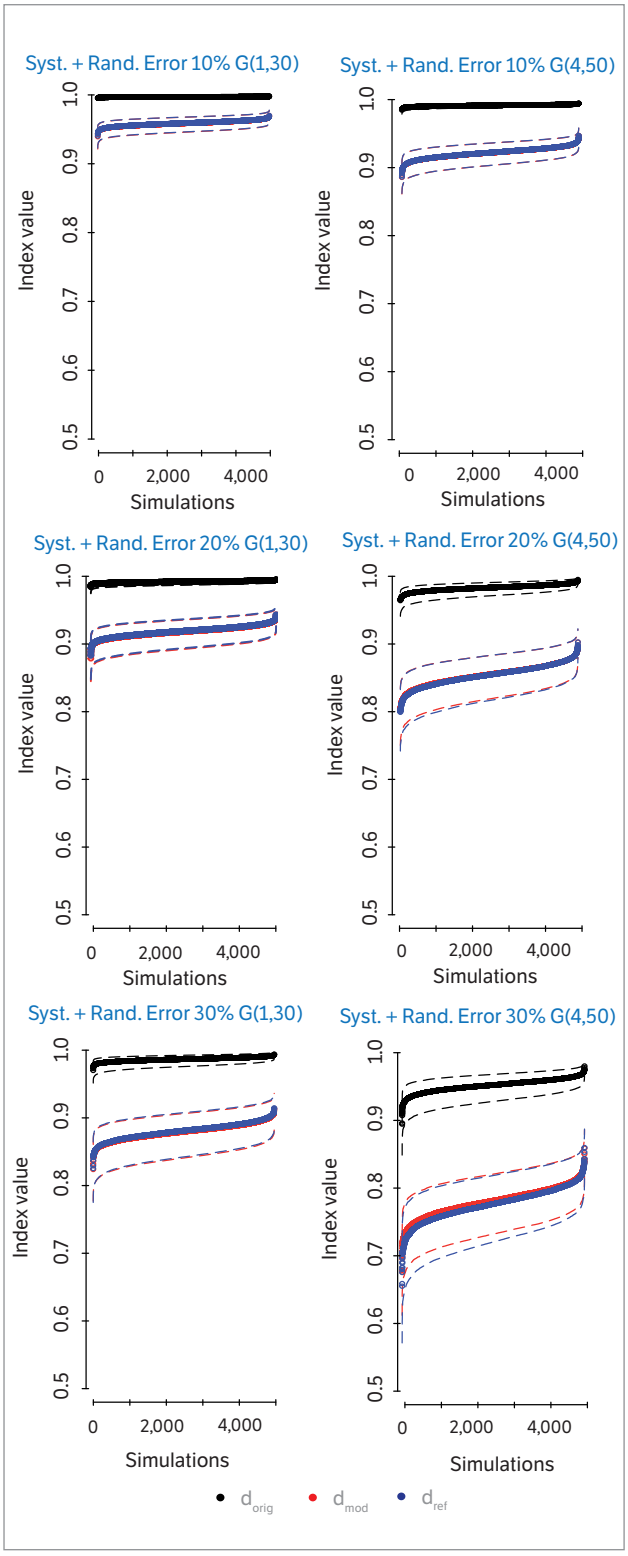

Figure 3. Performance of the original (d_{orig} black dots), modified (d_{mod} red dots) and refined (d_{ref} blue dots) indices of Agreement subjected to (positive) random+systematic errors. The observed values were generated from a 2-parameter gamma distribution [G(.)] with shape parameter set to 1 and 4 and scale parameter set to 30 and 50.

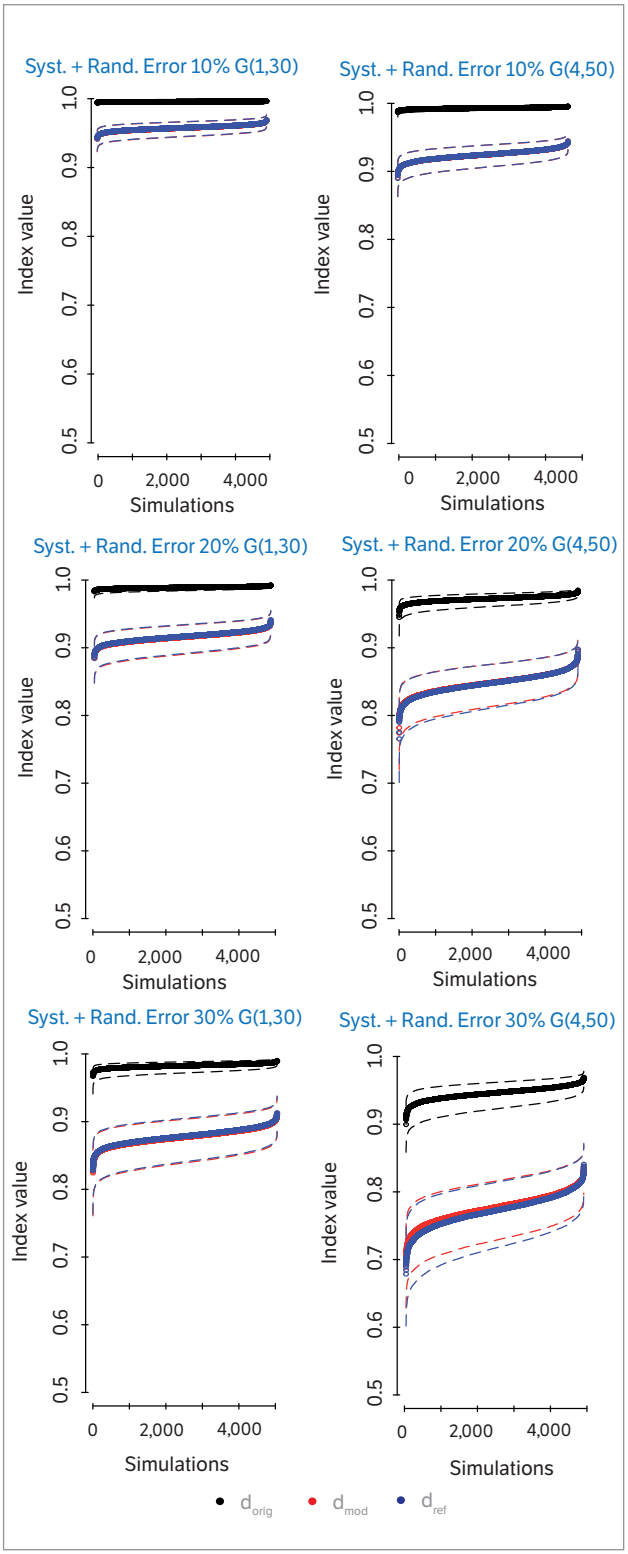

Figure 4. Performance of the original (d_{orig} black dots), modified (d_{mod} red dots) and refined (d_{ref} blue dots) indices of Agreement subjected to (negative) random+systematic errors. The observed values were generated from a 2-parameter gamma distribution [G(.)] with shape parameter set to 1 and 4 and scale parameter set to 30 and 50.

holds for all types of error and is exemplified in Figure 4. This is a natural behavior of indices developed to vary between zero and one, using the modulus and/or the square function in their calculation algorithm. On the other hand, considering that d_{ref} was developed to vary between –1.0 and 1.0, one could expect that errors with same magnitude but opposite signs would lead to different d_{ref} values. However, as demonstrated in Willmott et al. (2012), negative values of d_{ref} only indicate large predictive errors. This is the reason why Monte Carlo simulations revealed that the three indices presented similar behaviors in such cases (Figures 3 and 4). Therefore, as d_{mod} and d_{orig}, a given d_{ref} does not indicate that a predicting model overestimate or underestimate the simulated data.

The outcomes of the Monte Carlo Simulations also indicated that the three indices may assume different values for a particular errors magnitude (Figures 1 to 4) in respect to the process mean value. For instance, considering a gamma distribution with shape parameter equal to 1 and scale parameter equal to 30 [G(1,30)], d_{orig}, d_{mod} and d_{ref} respectively ranged from ~0.95 to ~098, ~0.78 to ~0.85 and ~0.78 to ~0.85 when the systematic errors were set to 20% (Figure 1). This feature implies that none of the three indices can be directly used to evaluate the average error of the predicted/estimated values in respect to the real/observed process average value. This statement is further supported by the fact that the three indices are affected by the parameters of the distribution from which the observed data were drawn. As previously described, d_{mod} ranged from ~0.78 to ~0.85 when the systematic error were set to 20% and the Gamma distribution G(1,30) were used to generate the observed data. Considering the same errors magnitude, d_{mod} ranged from ~0.63 to ~0.77 when G(4,50) were used to generate the observed data. This latter statement holds for the three indices and all sorts of errors.

Case Study

It is worth mentioning that atmospheric water demand is mainly driven by the following variables: incoming solar radiation, air temperature, wind speed and vapor pressure deficit (Vicente-Serrano et al. 2014; among many others). The PM equation has two components (energetic and aerodynamic) that seek to represent the contribution of all these variables to a given ETo value. In general, the energetic component contribution to an ETo daily amount is higher than the aerodynamic factor

contribution (Penman 1948[1]; Vicente-Serrano et al. 2014; Matsoukas et al. 2011). However, the significance of this latter component to a particular daily ETo value tends to increase in dry regions and/or during dry period/seasons (Matsoukas et al. 2011; McVicar et al. 2012). In other words, a relative decrease in the energetic component tends to be associated with an increase in the significance of the aerodynamic component. Finally, it is also worth emphasizing that the TW model considers only variables related to the energetic component of the atmospheric water demand, i.e., air temperature and those related to the photoperiod. On such theoretical background and considering the climate conditions of Campinas, the level of agreement between ETo daily amounts derived from both TW and PM models is expected to vary over the seasons, reaching its lower level during the austral winter (dry) season. Likewise, a particular index used to evaluate model performance is expected to indicate that the TW model cannot be applied to estimate ETo daily amounts in Campinas-SP.

The seasonal variability of the three indices calculated through the R-code is consistent with the above-mentioned theoretical background. In addition, the results depicted in Figure 5 also agree with those of the Monte Carlo simulations that indicated that both d_{ref} and d_{mod} are more rigorous than d_{orig}, and hence they should be preferred over their original version. This statement is particularly true for the summer season when, considering the 95% confidence interval, the results of Figure 5 indicated that d_{orig} may reach values as high as 0.80 in the presence of absolute mean errors larger than 0.80 mm·day^{-1}. Considering that d_{orig}=1 indicates a perfect model, the original version of the Willmott index may lead the user to erroneously accept that the TW model is (at least) suitable for estimating daily ETo amounts in Campinas during summer seasons. On the other hand, the upper limits (95% confidence interval) of all d_{mod} values remained lower than 0.60 in all seasons. Naturally, the d_{mod} values are consistent with the above-mentioned theoretical background, indicating →

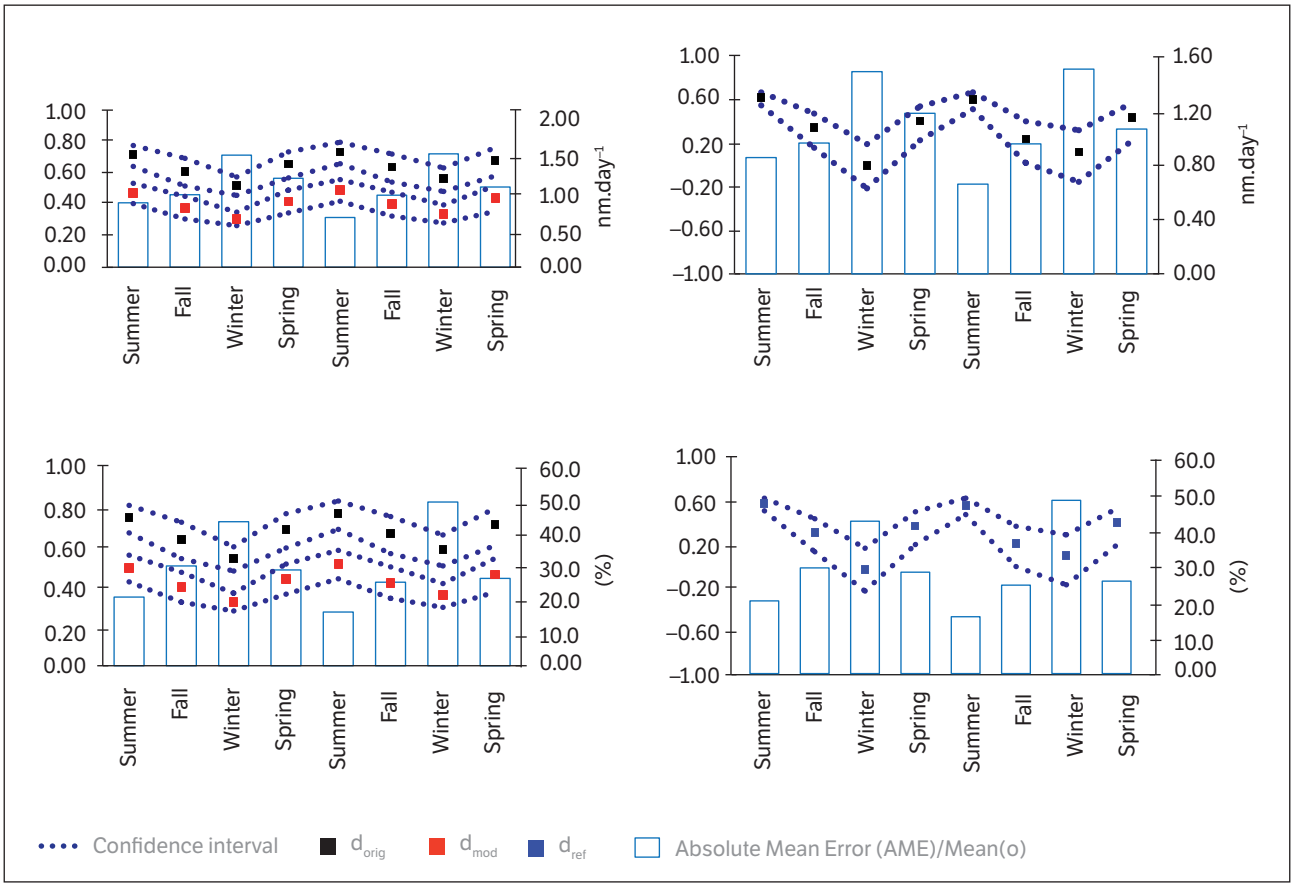

Figure 5. Original (d_{orig} black dots), modified (d_{mod} red dots) and refined (d_{ref} blue dots) indices of Agreement (left axis). The blue bars (right axis) represent the Absolute Mean Error (AME; a and b) and the ratio between AME and the mean value of the daily process [AME/Mean(o)]. The dashed lines represent the 95% confidence interval.

[1]Penman, H.L. (1948). Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society a Mathematical, Physical and Engineering Sciences, 193, 120-145. http://dx.doi.org/10.1098/rspa.1948.0037

that the TW model cannot be applied to estimate ETo daily amounts in Campinas-SP, regardless the season.

For the summer season, d_{ref} presented similar values as those shown by d_{mod}. However, d_{ref} assumed negative values in the winter of 2008. Considering that the Monte Carlo Simulations indicated that a particular d_{ref} does not inform the user if a predicting model overestimate or underestimate the simulated data, this negative value only indicates that the TW model had its worst performance in the winter of 2008. Finally, none of the three indices can be regarded as linear functions of AME/Mean(PM).

FINAL REMARKS

Considering the errors magnitude adopted in the Monte Carlo Simulations as well as the case of study, our findings indicate that the original version of the Willmott index may lead the user to erroneously select a predicting model that generates poor estimates. This statement is consistent with previous studies. Our results also indicate that the two newer versions of this index (modified and refined) overcome such problem, leading to more rigorous evaluations of the predicting models. Therefore, they should be preferred over the original version.

Although the refined Willmott index presents the broadest range of possible values [–1.0:1.0], it does not inform the user if a predicting model overestimate or underestimate the simulated data. Therefore, it added no extra information in respect to those already provided by the modified version of the agreement index. Naturally, this statement holds for the simulations and case of study carried out in this paper. None of the indices represents the error as linear functions of its magnitude in respect to the observed process.

ORCID IDs

H.R. Pereira
https://orcid.org/0000-0002-1927-9043
M.C. Meschiatti
https://orcid.org/0000-0002-9492-8672
R.C.M. Pires
https://orcid.org/0000-0003-4200-7094
G.C. Blain
https://orcid.org/0000-0001-8832-7734

REFERENCES

Allen, R. G., Pereira, L. S., Raes, D. and Smith, M. (1998). Crop evapotranspiration: guidelines for computing crop water requirements (FAO Irrigation and Drainage Paper, No. 56). Roma: FAO.

Blain, G. C. (2014). Revisiting the critical values of the Lilliefors test: towards the correct agrometeorological use of the Kolmogorov-Smirnov framework. Bragantia, 73, 192-202. http://dx.doi.org/10.1590/brag.2014.015.

Bardin-Camparotto, L., Blain, G. C., Giarolla, A., Adami, M. and Camargo, M. B. P. (2013). Validação de dados termo pluviométricos obtidos via sensoriamento remoto para o Estado de São Paulo. Revista Brasileira de Engenharia Agrícola e Ambiental, 17, 665-671. http://dx.doi.org/10.1590/S1415-43662013000600013.

Carvalho, L. M. V., Jones, C. and Liebmann, B. (2004). The South Atlantic Convergence Zone: Intensity, Form, Persistence, and Relationships with Intraseasonal to Interannual Activity and Extreme Rainfall. Journal of Climate, 17, 88-108. https://doi.org/10.1175/1520-0442(2004)017<0088:TSACZI>2.0.CO;2.

Carvalho, L. G., Rios, G. F. A., Miranda, W. L. and Castro Neto, P. (2011). Evapotranspiração de referência: uma abordagem atual de diferentes métodos de estimativa. Pesquisa Agropecuária Tropical, 41, 456-465. http://dx.doi.org/10.5216/pat.v41i3.12760.

Krause, P., Boyle, D. P. and Base, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5, 89-97. https://doi.org/10.5194/adgeo-5-89-2005.

Legates, D. R. and McCabe Jr., G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35, 233-241. https://doi.org/10.1029/1998WR900018.

Matsoukas, C., Benas, N., Hatzianastassiou, N., Paylakis, K.G., Kanakidou, M. and Vardavas, I. (2011). Potential evaporation trends over land between 1983-2008: driven by radiative fluxes or vapor-pressure deficit? Atmospheric Chemistry and Physics, 11, 7601-7616. https://doi.org/10.5194/acp-11-7601-2011.

McVicar, T. R., Roderick, M. L., Donohue, R. J., Li, L. T., Van Niel, T. G., Thomas, A., Grieser, J., Jhajharia, D., Himri, Y. and Mahowald, N. M. (2012). Global review and synthesis of trends in observed terrestrial near-surface wind speeds: implications for evaporation. Journal of Hydrology. 416, 182-205. https://doi.org/10.1016/j.jhydrol.2011.10.024.

Meschiatti, M. C. and Blain, G. C. (2016). Increasing the regional availability of the Standardized Precipitation Index: an operational approach. Bragantia, 75, 507-521. http://dx.doi.org/10.1590/1678-4499.478.

Vicente-Serrano, S. M., Azorin-Molina, C., Sanchez-Lorenzo, A., Revuelto, J., López-Moreno, J. I., González-Hidalgo, J. C., Moran-Tejeda, E. and Espejo, F. (2014). Reference evapotranspiration variability and trends in Spain, 1961-2011. Glob Planet Change, 121, 26-40. https://doi.org/10.1016/j.gloplacha.2014.06.005.

Wilks, D. S. (2011). Statistical methods in the atmospheric sciences. San Diego: Academic Press.

Willmott, C. J. (1981). On the validation of models. Physical Geography, 2, 184-194. http://doi.org10.1080/02723646.1981.10642213.

Willmott, C. J. (1982). Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society, 63, 1309-1313. https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2.

Willmott, C. J. (1984). On the evaluation of model performance in physical geography. In G. L. Gaile and C. J. Willmott (Eds.). Spatial Statistics and Models, (p.443-460). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-3048-8_23.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., O'donnell, J. and Rowe, C. M. (1985). Statistics for the evaluation of model performance. Journal of Geophysical Research, 90, 8995-9005. https://dx.doi.org/10.1029/JC090iC05p08995.

Willmott, C. J., Matsuura, K. and Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. Atmospheric Environment, 43, 749-752. https://doi.org/10.1016/j.atmosenv.2008.10.005.

Willmott, C. J., Robeson, S. M. and Matsuura, K. A. (2012). A refined index of model performance. International Journal of Climatology, 32, 2088-2094. https://doi.org/10.1002/joc.2419.

Willmott, C. J. and Wicks, D. E. (1980). An Empirical method for the spatial interpolation of monthly precipitation within California. Physical Geography, 1, 59-73. http://doi.org10.1080/02723646.1980.10642189?journalCode=tphy20.

Wu, H., Hayes, M. J., Wilhite, D. A. and Svoboda, M. D. (2005). The effect of the length of record on the Standardized Precipitation Index calculation. International Journal of Climatology, 25, 505-520. https://doi.org/10.1002/joc.1142.