Environmental Modelling & Software 104 (2018) 236–248

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A modified F-test for evaluating model performance by including both experimental and simulation uncertainties

Nathan Q. Sima [a], R. Daren Harmel [b], Quan X. Fang [c], Liwang Ma [d], Allan A. Andales [a, *]

[a] Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA
[b] Center for Agricultural Resources Research Center, USDA-ARS, Fort Collins, CO, USA
[c] State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Soil and Water Conservation, Chinese Academy of Sciences, Yangling, 712100, China
[d] Rangeland Resources and Systems Research Unit, USDA-ARS, Fort Collins, CO, USA

ARTICLE INFO

Article history:
Received 5 October 2017
Received in revised form
27 January 2018
Accepted 15 March 2018

Keywords:
Uncertainty
Model calibration
RZWQM
F-test
Goodness of fit
Maize

ABSTRACT

Experimental and simulation uncertainties have not been included in many of the statistics used in assessing agricultural model performance. The objectives of this study were to develop an F-test that can be used to evaluate model performance considering experimental and simulation uncertainties, and identify the best datasets to use for model calibration using different water stress functions in a cropping system model. Data on irrigated maize in Colorado, USA, and the Root Zone Water Quality Model (RZWQM) were used as an example to demonstrate model calibration using the modified F-test along with other commonly used statistics. Compared to the d-index, the F-test provided a statistical test under a certain confidence level that better distinguished the goodness of model prediction for both biomass and yield while considering uncertainty. To obtain robust model parameters, we recommend using multiple treatments across multiple years for model calibration, regardless of water stress functions used.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Very few statistics are satisfactory in determining the goodness of model fit in calibration and prediction, and model users seem to tolerate a large simulation error in agriculture because of considerable experimental uncertainties in field research (Ritter and Muñoz-Carpena, 2013). The most commonly used statistical metrics for agricultural models are coefficient of determination (r^2), root mean squared error (RMSE), relative error (RE), Nash-Sutcliffe model efficiency (NSE), and index of agreement (d-index), in addition to graphical visualization (Ma et al., 2011; Krause et al., 2005; Whitmore, 1991). However, these statistics do not consider either experimental and simulation uncertainties. Therefore, it may be erroneous or subjective to use these statistics in judging the performance of a model using these statistics based on the criteria suggested by Moriasi et al. (2007). Ritter and Muñoz-Carpena (2013) used bootstrapping to create empirical probability distributions of RMSE and NSE so that significance of RMSE and NSE in relation to their respective threshold could be tested, but they still did not consider experimental nor simulation uncertainties. In addition, the significance was referring to RMSE and NSE, not to the experimental nor simulation data.

To account for experimental (measurement) uncertainties, Harmel and Smith (2007) introduced a correction factor (CF) for calculating the differences between simulated and observed values in their statistics. They found that model performance (d-index) was improved after considering experimental uncertainty. In a later study, Harmel et al. (2010) further improved the d-index by incorporating both experimental and simulation uncertainties. Harmel et al. (2014) stated that it was critically important to include both measured data and simulation uncertainty in interpreting and communicating model results. Yen et al. (2014) developed an integrated parameter estimation and uncertainty analysis tool (IPEAT) that considered uncertainties due to input data, model parameters, model structure, and calibration/validation data, using the correction factor (CF) of Harmel and Smith (2007) for the Nash-Sutcliffe model efficiency. They found that the calibrated model, by considering three or more uncertainties, was more robust than that without considering any uncertainty in model parameterization. However, in a later study, Yen et al. (2016) found that including

* Corresponding author. 1170 Campus Delivery, Fort Collins, CO, 80523-1170, USA.
E-mail address: Allan.Andales@ColoState.edu (A.A. Andales).

https://doi.org/10.1016/j.envsoft.2018.03.011
1364-8152/© 2018 Elsevier Ltd. All rights reserved.

measurement uncertainty in calibration datasets would not significantly affect model calibration unless the uncertainty is greater than 50%.

However, most statistics used in model comparisons are relative and there is no test of significance between model performances. For example, Saseendran et al. (2014) compared three water stress functions in the Root Zone Water Quality Model (RZWQM) using the RMSE and d-index. They found that the two modified water stress functions based on the Nimah-Hanks water uptake or canopy heating effects provided better predictions of maize yield, biomass, leaf area index, and soil water content using data from three locations in Colorado than the original water uptake function in the CERES-Maize model. In another study, Fensterseifer et al. (2017) used relative RMSE as a criterion to evaluate the number of datasets needed to calibrate the CROPGRO-soybean model using 21 datasets from eight locations in Southern Brazil. They found that two datasets from each location are needed for model calibration. However, both studies used trial-and-error in model calibration, and the calibrated parameters may not be rigorous in addition to the subjectivity of the statistics used. Ma et al. (2012a) and Kersebaum et al. (2008) used a Lack-of-Fit (an F-test) to calculate the significance of model predictions when experimental uncertainties existed, based on the study of Whitmore (1991). They found that the F-test could be more rigorous than the ranges of indicators proposed by Moriasi et al. (2007). The objectives of this study were to: (1) develop an F-test for evaluating model performance by including both experimental and simulation uncertainties, (2) compare model performance using various water stress functions in RZWQM, and (3) identify the best datasets for model calibration. A published dataset on irrigated maize in Colorado, USA, and the Root Zone Water Quality Model (RZWQM) were used as an example to demonstrate the modified F-test along with three commonly used statistics: the d-index, RMSE, and r^2. Since the modeling results were published in previous papers (Ma et al., 2012b, 2016; Saseendran et al., 2014), our emphasis was on quantifying the model performance by applying the newly modified F-test, rather than analyzing the causes of modeling deficiency from the underlying biophysical processes in the model.

2. Materials and methods

2.1. Experimental dataset and model simulation

The field experimental data were obtained from a study conducted from 2008 to 2011 near Greeley, Colorado, USA (40.45° N, 104.64° W). The soil is a sandy loam and is fairly uniform throughout the 200 cm soil profile. Six irrigation treatments (micro-irrigation with surface drip tubing adjacent to each row) with four replicates were designed to meet a specified percentage of potential crop evapotranspiration (ET) requirements (Allen et al., 1998, 2005) during the growing seasons: 100% (T1), 85% (T2), 75% (T3), 70% (T4), 55% (T5) and 40% (T6) of potential crop ET. The amount of water for each treatment was estimated at a 3–6 day interval based on reference ET demand, crop coefficient, rainfall, and soil water deficit. The T1 treatment was irrigated such that water availability (irrigation plus precipitation plus stored soil water) was adequate to meet crop water requirements, as predicted by the reference evapotranspiration and crop coefficients (FAO-56 methodology, Allen et al., 1998). The remaining treatments were irrigated to meet a certain percentage of water demand in T1.

Maize cv. Dekalb 52–59 was planted at an average rate of 81,000 seeds per hectare with 0.76 m row spacing in early May from 2008 to 2011. A detailed description of the experiment is provided by Ma et al. (2012b), and the experimental dataset and detailed methodology can also be found at US Department of Agriculture National Agricultural Library Ag Data Commons (Trout and Bausch, 2017).

The Root Zone Water Quality Model (RZWQM) is a comprehensive agricultural system model and has process-level simulations of soil water, soil temperature, plant growth, pesticide fate, and soil C and N dynamics as influenced by various agricultural management practices (Ahuja et al., 2000). The DSSAT 4.0 crop models (e.g., CERES-Maize and CERES-Wheat models) incorporated in RZWQM can be used to simulate crop growth, water use, and N uptake, where RZWQM provides soil water, soil temperature, and nutrient information for DSSAT4.0 crop models (Ma et al., 2006).

In RZWQM, there are three water stress functions users can select. The first is from the DSSAT model that is defined as the ratio of potential root water uptake (*TRWUP*) to potential plant transpiration (EP_o) (Ritchie, 1998), referred hereafter as default water stress function (WSF1, Saseendran et al., 2014).

$$WSF1 = \frac{TRWUP}{EP_O} \tag{1}$$

The simplified close-form equation of (Ritchie, 1998) used to calculate the *TRWUP* in Eq. (1) is:

$$\mathrm{TRWUP} = \sum_{i=1}^{N} \frac{k1 * e^{k2*(SW(i)-LL(i))}}{k3 - \ln(RLV(i))} * RLV(i) * \Delta Z(i) \tag{2}$$

where, *RLV(i)* is root length density in soil layer *i* (cm cm^{-3}); *k1* = 0.00132, *k2* = 45.0 if the drained lower limit (LL) of soil water (permanent wilting point or soil water content at 1.5 MPa suction) in the soil layer is greater than 0.30 cm^3 cm^{-3} and *k2* = 130 LL(L), if LL for the soil layer is less than 0.30 cm^3 cm^{-3}; *k3* = 7.01; *SW(i)* and *LL(i)* are, respectively, volumetric soil water content and lower limit of plant available water in layer *i* (cm cm^{-1}); *Z(i)* is soil depth of layer *i* (cm).

The second water stress function (WSF2) estimates TRWUP from the Nimah-Hanks equation (Nimah and Hanks, 1973). The root water uptake part of the sink term, S_r *(z, t)* (cm hr^{-1}), is computed using the Nimah and Hanks (1973) equation:

$$TRWUP_{NH} = \sum_{i=1}^{N} S_r(z_i, t) = \sum_{i=1}^{N} \frac{[H_r + (R_r z_i) - h(z_i, t) - s(z_i, t)]R(z_i)K_i(\theta)}{\Delta x \Delta z_i} \Delta z_i \tag{3}$$

where, θ = volumetric soil water content (cm^3 cm^{-3}); t = time (hr); z_i = soil depth (cm, assumed positive downward); h = soil-water pressure head (cm); $K_i(\theta)$ = unsaturated hydraulic conductivity (cm hr^{-1}), a function of h and z; H_r = an effective root water pressure head (cm); R_r = a root resistance term and the product (R_r z_i) accounts for gravity term and friction loss in H_r (assumed = 1.05); $s(z_i,t)$ = the osmotic pressure head (assumed = 0 cm); Δx = the distance from plant roots to where $h(z_i,t)$ is measured (assumed = 1 cm); Δz = soil depth increment (cm); $R(z_i)$ = proportion of the total root activity in the depth increment Δz_i, obtained from the plant growth model. The total potential uptake (*TRWUP*$_{NH}$) is calculated from summation of Eq. (3) with H_r set equal to −1.5 MPa as the permanent wilting point, which may be modified by users according to crop species. Thus,

$$WSF2 = \frac{TRWUP_{NH}}{EP_O} \tag{4}$$

The third water stress function considers water stress due to heating of the canopy by the latent heat energy partitioned to

potential soil evaporation but not used in soil evaporation when the surface soil water content is limiting. Therefore, we explored including stress due to additional canopy heating in calculation of the water stress functions by changing their formulation as described below by including actual soil evaporation (E_S) for the day in the numerator and using total evapotranspiration (ET) in the denominator.

$$WSF3 = \frac{TRWUP_{NH} + E_S}{ET} \tag{5}$$

With measured soil properties shown in Table 1, an automated optimization and parameter estimation software (PEST, Doherty, 2010; Ma et al., 2012a, 2016) in RZWQM was used to calibrate crop cultivar parameters (Tables 2 and 3). The six irrigation treatments with four replicates carried out from 2008 to 2011 were used for selecting subsets of observed data to be used for model calibration. To investigate the effect of different measured datasets on optimizing crop cultivar parameters and crop growth outputs, RZWQM was first calibrated either using sub-datasets from one treatment of multiple years (Trt-1, Trt-2, Trt-3, Trt-4, Trt-5, and Trt-6) or from multiple treatments of one year (Year-2008, Year-2009, Year-2010, and Year-2011), or all the treatments and the years. The calibrated cultivar parameters were then used to simulate all the datasets and statistics were computed from all the datasets simulated with each set of cultivar parameters. Detailed information on model calibration using PEST in RZWQM can be found in Ma et al. (2012a).

The three water stress functions have been modeled and compared with the same data above by Saseendran et al. (2014) and we are using it again here to re-evaluate the performance of the WSFs based on the modified F-test below. Different from Saseendran et al. (2014), the PEST was implemented to derive crop parameters, rather than by trial-and-error.

2.2. Modified F-test with experimental and simulation uncertainty

For convenience, we first define the following symbols:

L = number of experimental or measurement groups. The groups may represent different treatments or different sampling dates.
N_i = number of measured replicates for the i^{th} experimental or measurement group
M_i = number of prediction replicates for the i^{th} experimental or measurement group
O_{ij} = j^{th} observation (replicate) for the i^{th} measurement group ($O_{ij} = \mu_i + \varepsilon_{ij}$ and $E[O_{ij}] = \mu_i$)
P_{ik} = k^{th} predicted value (replicate) for the i^{th} measurement group ($P_{ik} = \lambda_i + \delta_{ik}$ and $E[P_{ik}] = \lambda_i$)
λ_i = true mean of predictions for the i^{th} experimental or measurement group
μ_i = true mean of observations for the i^{th} experimental or measurement group
$P_i = \frac{1}{M_i}\sum_{k=1}^{M_i} P_{ik}$ = mean of prediction for the i^{th} experimental or measurement group based on a simulation model. P_i is independent of O_{ij}.
$O_i = \frac{1}{N_i}\sum_{j=1}^{N_i} O_{ij}$ = mean of the i^{th} experimental or measurement group
$D_i^2 = \frac{1}{M_i-1}\sum_{k=1}^{M_i}(P_{ik} - P_i)^2$ = prediction variance of i^{th} experimental or measurement group
$S_i^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(O_{ij} - O_i)^2$ = sample variance of i^{th} experimental or measurement group

When there is no simulation uncertainty, for an experiment with N experimental or measurement groups and K_i replicates in each group, total sum of squared prediction errors (TSS) may be written as (Ma et al., 2012a):

$$TSS = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(P_i - O_{ij})^2 \tag{6}$$

which can be rearranged as:

$$\begin{aligned} TSS &= \sum_{i=1}^{L}\sum_{j=1}^{N_i}\left[(P_i - O_i) + (O_i - O_{ij})\right]^2 \\ &= \sum_{i=1}^{L}\sum_{j=1}^{N_i}(P_i - O_i)^2 + \sum_{i=1}^{L}\sum_{j=1}^{N_i}(O_{ij} - O_i)^2 \\ &= \sum_{i=1}^{L} N_i(P_i - O_i)^2 + \sum_{i=1}^{L}\sum_{j=1}^{N_i}(O_{ij} - O_i)^2 \\ &= LOFIT + SSE \end{aligned} \tag{7}$$

Where LOFIT is the sum of squared errors between predicted and observed mean values (due to lack of fit) and SSE is the sum of squared error due to experimental error (ε_{ij}). SSE may be rewritten as (Wackerly et al., 2008):

$$LOFIT = \sum_{i=1}^{L} N_i(P_i - O_i)^2 \tag{8}$$

and

$$SSE = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(O_{ij} - O_i)^2 = \sum_{i=1}^{L}(N_i - 1)S_i^2 \tag{9}$$

The mean LOFIT (MSLOFIT) and mean SSE (MSE) are defined as:

$$MSLOFIT = \frac{LOFIT}{\sum_{i=1}^{L} N_i}, \quad and\ MSE = \frac{SSE}{\sum_{i=1}^{L}(N_i - 1)} = \frac{\sum_{i=1}^{L}(N_i - 1)S_i^2}{\sum_{i=1}^{L}(N_i - 1)} \tag{10}$$

Table 1
Soil parameters and range of cultivar parameters for PEST optimization (Ma et al., 2016).

Soil depth (cm)	Soil bulk density (g/cm^3)	Saturated soil water content (cm^3/cm^3)	Measured averaged field capacity (cm^3/cm^3)
0–15	1.492	0.437	0.258
15–45	1.492	0.437	0.239
45–75	1.492	0.437	0.211
75–105	1.568	0.408	0.185
105–135	1.568	0.408	0.182
130–165	1.617	0.390	0.183
160–190	1.617	0.390	0.209

Table 2
Optimized crop parameters fitted for one treatment across four years. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation.

Parameter Name and (Ranges, initial values) for PEST optimization	WSF	CV	Parameter values fitted from each treatment across 4 years						
			Trt-1	Trt-2	Trt-3	Trt-4	Trt-5	Trt-6	Trt-All
P1 - Degree days (base temperature of 8 °C) from seedling emergence to end of juvenile phase (thermal degree days) (100–450, 250).	WSF1	0.081	237.5	277.3	262.8	239.4	225.9	227.0	263.9
	WSF2	0.027	240.6	250.5	253.6	255.1	248.3	261.8	247.2
	WSF3	0.020	260.7	259.8	250.3	251.9	248.7	252.3	248.2
P2 - Day length sensitivity coefficient [the extent (days) that development is delayed for each hour increase in photoperiod above the longest photoperiod (12.5 h) at which development proceeds at maximum rate] (0–1, 0.2).	WSF1	0.981	0.19	0.01	0.54	0.11	0.179	0.86	0.22
	WSF2	0.954	0.21	0.19	0.21	0.16	0.07	0.83	0.19
	WSF3	0.524	0.04	0.23	0.19	0.18	0.23	0.04	0.20
P5 - Degree days (base temperature of 8 °C) from silking to physiological maturity (thermal degree days) (500–1000, 600)	WSF1	0.153	697.8	688.2	717.7	974.0	755.2	611.3	725.0
	WSF2	0.122	757.4	701.9	684.4	725.8	671.9	500.0	681.3
	WSF3	0.066	692.9	683.2	696.9	699.3	672.9	573.6	688.5
G2 - Potential kernel number per plant (440–1000, 900)	WSF1	0.298	972.8	1000.0	598.5	1000.0	922.0	472.6	554.3
	WSF2	0.227	953.0	991.8	963.8	885.2	762.8	451.8	975.4
	WSF3	0.073	979.8	946.5	967.4	803.0	1000.0	1000.0	924.3
G3 - Potential kernel growth rate (mg/(kernel d) (5–16, 6)	WSF1	0.374	6.45	6.72	10.10	5.42	7.66	15.01	10.42
	WSF2	0.456	6.05	6.34	6.34	6.25	7.14	15.71	6.35
	WSF3	0.081	6.32	6.82	6.55	7.44	5.69	6.53	6.80
PHINT - Degree days required for a leaf tip to emerge (thermal degree days) (38–55, 50)	WSF1	0.105	48.1	45.9	44.7	38.0	50.3	38.5	42.5
	WSF2	0.035	49.4	48.5	51.8	50.8	50.5	53.9	51.6
	WSF3	0.028	49.3	51.1	50.1	48.7	52.8	51.9	51.0

Therefore, an F-test statistic can be constructed as (Whitmore, 1991; Kersebaum et al., 2008):

$$F1 = \frac{MSLOFIT}{MSE} \tag{11}$$

with degrees of freedom of $\nu_1 = \sum_{i=1}^{L} N_i$ for the numerator and $\nu_2 = \sum_{i=1}^{L} (N_i - 1)$ for the denominator. To test whether model predictions P_i correctly estimate the true mean of the observations for the ith experimental or measurement group, the hypothesis would be:

H_o: $P_i = \mu_i$ for all *i*
H_a: $P_i \neq \mu_i$ for at least one *i*

The rejection of H_o would indicate a 'lack of fit' of the simulations with respect to the true experimental means. At a given level of significance (e.g., α level), a critical $F_{\alpha,\nu1,\nu2}$ value can be used to test the acceptability of the null hypothesis. In this study, $\nu_1 = 96$ and $\nu_2 = 72$ as $N_i = 4$ and $L = 24$.

When there is both experimental and simulation uncertainties, for an experiment with L experimental or measurement groups, N_i measurement replicates in each group, and M_i predictions in each group, total sum of squared prediction errors (TSS) may be written as:

$$TSS = \sum_{i=1}^{L} \sum_{j=1}^{N_i} \sum_{k=1}^{M_i} (P_{ik} - O_{ij})^2 \tag{12}$$

Table 3
Optimized crop parameters fitted for all treatments in one year. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation.

Parameter Name and (Ranges, initial values) for PEST optimization	WSF	CV	Parameters fitted from all treatments in each year			
			Year-2008	Year-2009	Year-2010	Year-2011
P1 - Degree days (base temperature of 8 °C) from seedling emergence to end of juvenile phase (thermal degree days). (100–450, 250)	WSF1	0.122	307.8	236.8	244.4	275.0
	WSF2	0.143	255.6	257.5	203.2	197.3
	WSF3	0.155	275.7	253.0	190.3	226.8
P2 - Day length sensitivity coefficient [the extent (days) that development is delayed for each hour increase in photoperiod above the longest photoperiod (12.5 h) at which development proceeds at maximum rate]. (0–1, 0.2)	WSF1	0.339	0.38	0.26	0.27	0.52
	WSF2	0.728	0.22	0.19	0.02	0.38
	WSF3	0.741	0.17	0.20	0.10	0.51
P5 - Degree days (base temperature of 8 °C) from silking to physiological maturity (thermal degree days) (500–1000, 600)	WSF1	0.291	1000.0	591.8	568.6	629.7
	WSF2	0.206	650.6	537.9	842.7	831.7
	WSF3	0.089	645.5	591.2	707.8	587.8
G2 - Potential kernel number per plant (440–1000, 900)	WSF1	0.412	459.9	971.8	470.5	989.8
	WSF2	0.019	984.6	954.6	984.4	1000.0
	WSF3	0.106	765.4	942.8	751.6	820.2
G3 - Potential kernel growth rate (mg/(kernel d) (5–16, 6)	WSF1	0.365	13.31	6.08	16.00	16.00
	WSF2	0.114	5.90	6.35	5.05	5.09
	WSF3	0.227	8.22	6.37	8.04	10.97
PHINT - Degree days required for a leaf tip to emerge (thermal degree days) (38–55, 50)	WSF1	0.187	39.5	39.6	55.0	38.0
	WSF2	0.112	43.9	50.4	55.0	44.0
	WSF3	0.069	48.0	47.9	55.0	52.7

Which can be rearranged as:

$$\begin{aligned} TSS &= \sum_{i=1}^{L}\sum_{j=1}^{N_i}\sum_{k=1}^{M_i}\left[(P_{ik}-P_i)+(P_i-O_i)+(O_i-O_{ij})\right]^2 \\ &= \sum_{i=1}^{L}\sum_{j=1}^{N_i}\sum_{k=1}^{M_i}(P_i-O_i)^2+\sum_{i=1}^{L}\sum_{j=1}^{N_i}\sum_{k=1}^{M_i}(P_{ik}-P_i)^2+\sum_{i=1}^{L}\sum_{j=1}^{N_i}\sum_{k=1}^{M_i}(O_i-O_{ij})^2 \\ &= \sum_{i=1}^{L}M_iN_i(P_i-O_i)^2+\sum_{i=1}^{L}N_i\sum_{k=1}^{M_i}(P_{ik}-P_i)^2+\sum_{i=1}^{L}M_i\sum_{j=1}^{N_i}(O_i-O_{ij})^2 \\ &= LOFIT+SSP+SSE \end{aligned} \tag{13}$$

Where LOFIT is the sum of squared errors between predicted and observed mean values (due to lack of fit), SSE is the sum of squared error due to experimental uncertainty (ε_{ik}), and SSP is the sum of squared error due to prediction uncertainty (δ_{ij}). SSP and SSE may be rewritten as (Wackerly et al., 2008):

$$SSP = \sum_{i=1}^{L}N_i\sum_{k=1}^{M_i}(P_{ik}-P_i)^2 = \sum_{i=1}^{L}N_i(M_i-1)D_i^2 \tag{14}$$

$$SSE = \sum_{i=1}^{L}M_i\sum_{j=1}^{N_i}(O_i-O_{ij})^2 = \sum_{i=1}^{L}M_i(N_i-1)S_i^2 \tag{15}$$

The mean LOFIT (MSLOFIT) and mean (SSP+SSE) (MSPE) are defined as:

$$MSLOFIT = \frac{LOFIT}{\sum_{i=1}^{L}M_iN_i},$$

$$\text{and} \tag{16}$$

$$MSPE = \frac{\sum_{i=1}^{L}N_i(M_i-1)D_i^2}{\sum_{i=1}^{L}N_i(M_i-1)}+\frac{\sum_{i=1}^{L}M_i(N_i-1)S_i^2}{\sum_{i=1}^{L}M_i(N_i-1)}$$

Therefore, an F-test statistic can be constructed as (Whitmore, 1991; Kersebaum et al., 2008):

$$F2 = \frac{MSLOFIT}{MSPE} \tag{17}$$

with degrees of freedom of $\nu_1 = \sum_{i=1}^{L}M_iN_i$ for the numerator and $\nu_2 = \sum_{i=1}^{L}N_i(M_i-1)+\sum_{i=1}^{L}M_i(N_i-1)$ for the denominator. To test whether model predictions λ_i correctly estimate the true mean of the observations for the ith experimental or measurement group, the hypothesis would be:

H_o: $\lambda_i = \mu_i$ for all i
H_a: $\lambda_i \neq \mu_i$ for at least one i

The rejection of H_o would indicate a 'lack of fit' of the simulations with respect to the true experimental means. At a given level of significance (e.g., α level), a critical $F_{\alpha,\nu1,\nu2}$ value can be used to test the acceptability of the null hypothesis. If the calculated F value does not exceed the critical F value, the null hypothesis is accepted. Otherwise, the null hypothesis is rejected, indicating a 'Lack of Fit' of the simulation results to the observed means. Since the critical value of $F_{\alpha,\nu1,\nu2}$ increases with decreasing α level at the same degrees of freedoms ($\nu_1 = 384$ and $\nu_2 = 576$ when $M_i = N_i = 4$ and $L = 24$ in this study), the null hypothesis may be rejected more easily at high α level.

2.3. Modified d-index considering experimental and simulation uncertainty

Willmott (1981) introduced an index of agreement (d-index):

$$d = 1 - \frac{\sum_{i=1}^{L}(O_i-P_i)^2}{\sum_{i=1}^{L}\left(|P_i-O_{avg}|+|O_i-O_{avg}|\right)^2} \tag{18}$$

To take into account the uncertainty of measured data used for calibration and evaluation, Harmel and Smith (2007) introduced a correction factor (CF) to modify the numerator in the above equation:

$$d1 = 1 - \frac{\sum_{i=1}^{L}\left[\frac{CF(meas)_i}{0.5}(O_i-P_i)\right]^2}{\sum_{i=1}^{L}\left(|P_i-O_{avg}|+|O_i-O_{avg}|\right)^2} \tag{19}$$

For a normal distribution of experimental uncertainty, $CF(meas)_i$ ranges from 0 for $O_i = P_i$ to 0.5 when P_i is 3.9 standard deviation away from O_i (Harmel and Smith, 2007).

When there are uncertainties in both experimental and simulation results, a new CF function is introduced (Harmel et al., 2010) and a d2-index is defined as:

$$d2 = 1 - \frac{\sum_{i=1}^{L}\left[CF(meas+pred)_i(O_i-P_i)\right]^2}{\sum_{i=1}^{L}\left(|P_i-O_{avg}|+|O_i-O_{avg}|\right)^2} \tag{20}$$

where $CF(meas+pred)_i = 1 - DO_i$, DO_i is the degree of overlap for distributions for each measured (O_i) and predicted (P_i) pair (Harmel et al., 2010), and O_{avg} is the mean of measured values.

2.4. Other statistics

In addition to the modified F-test and d-index, two other most commonly used simple statistics are relative Root Mean Squared Error (RRMSE) and coefficient of determination (r^2), which are:

Table 4
F-test, d-index, RRMSE (%), and r^2 for predicting biomass for the three water stress functions with simulation CV of 0.075 for one treatment across four years. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation. For each treatment (Trt-1 to Trt-6), L = 4 and N_i = 4; for Trt-All, L = 24, N_i = 4.

WSF	Calibration Treatment	d-index	r^2	RRMSE	Experimental Uncertainty only			Experimental and Simulation Uncertainty		
					d1-index	F1-value	p-value	d2-index	F2-value	p-value
WSF1	Trt-1	0.738	0.915	11.43	0.782	1.848	0.0034	0.911	0.961	0.66
	Trt-2	0.813	0.876	8.49	0.855	1.020	0.47	0.957	0.492	1.00
	Trt-3	0.824	0.884	8.00	0.864	0.906	0.67	0.965	0.432	1.00
	Trt-4	0.805	0.866	10.20	0.842	1.472	0.043	0.935	0.722	1.00
	Trt-5	0.801	0.926	9.57	0.844	1.296	0.12	0.939	0.656	1.00
	Trt-6	0.864	0.903	7.07	0.890	0.708	0.94	0.978	0.335	1.00
	Trt-All	0.846	0.911	7.05	0.889	0.704	0.95	0.971	0.338	1.00
WSF2	Trt-1	0.764	0.872	9.42	0.808	1.256	0.15	0.941	0.588	1.00
	Trt-2	0.755	0.874	10.23	0.792	1.480	0.041	0.928	0.683	1.00
	Trt-3	0.750	0.860	10.21	0.794	1.474	0.042	0.920	0.690	1.00
	Trt-4	0.735	0.835	10.54	0.777	1.572	0.022	0.919	0.736	1.00
	Trt-5	0.776	0.914	8.90	0.821	1.122	0.305	0.940	0.548	1.00
	Trt-6	0.712	0.802	11.44	0.755	1.851	0.0033	0.891	0.885	0.90
	Trt-All	0.781	0.893	8.91	0.825	1.122	0.305	0.941	0.534	1.00
WSF3	Trt-1	0.751	0.826	10.27	0.795	1.494	0.037	0.935	0.690	1.00
	Trt-2	0.748	0.831	10.59	0.790	1.587	0.020	0.927	0.731	1.00
	Trt-3	0.768	0.856	9.56	0.816	1.293	0.13	0.939	0.603	1.00
	Trt-4	0.784	0.883	8.88	0.830	1.115	0.31	0.943	0.526	1.00
	Trt-5	0.794	0.911	8.54	0.839	1.031	0.44	0.944	0.509	1.00
	Trt-6	0.729	0.861	10.47	0.781	1.550	0.026	0.912	0.768	1.00
	Trt-All	0.795	0.884	8.46	0.841	1.011	0.48	0.949	0.481	1.00

$$RRMSE = \frac{\sqrt{\frac{1}{L}\sum_{i=1}^{L}(P_i - O_i)^2}}{\frac{1}{L}\sum_{i=1}^{L} O_i} \quad (21)$$

$$r^2 = \frac{\left[\sum_{i=1}^{L}(O_i - O_{avg})(P_i - P_{avg})\right]^2}{\sum_{i=1}^{L}(O_i - O_{avg})^2 \sum_{i=1}^{L}(P_i - P_{avg})^2} \quad (22)$$

Where $O_{avg} = \frac{1}{L}\sum_{i=1}^{L} O_i$ and $P_{avg} = \frac{1}{L}\sum_{i=1}^{L} P_i$

Among these statistics, only the F-test is a statistical test for significance. The d-index, RRMSE, and r^2 are only numerical values and used for relative comparison among various model runs. Their implication for model performance is relative and is subject to interpretation. We selected the d-index to compare with the modified F-test because of the inclusion of experimental and simulation uncertainty as developed by Harmel and Smith (2007) and Harmel et al. (2010). We included RRMSE to measure the distance between simulated and measured results and r^2 to quantify the correlation between simulated and measured data. Both are basic statistics used by all scientific disciplines. To quantify the uncertainty of either measured or simulated results, we used the coefficients of variation (CV), also known as relative standard deviation (standard deviation/mean of a distribution), which measures the dispersion of a distribution of a variable.

Table 5
F-test, d-index, RRMSE (%), and r^2 for predicting yield for the three water stress functions with simulation CV of 0.075 for one treatment across four years. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation. For each treatment (Trt-1 to Trt-6), L = 4 and N_i = 4; for Trt-All, L = 24, N_i = 4.

WSF	Calibration Treatment	d-index	r^2	RRMSE	Experimental Uncertainty only			Experimental and Simulation Uncertainty		
					d1-index	F1-value	p-value	d2-index	F2-value	p-value
WSF1	Trt-1	0.755	0.849	13.52	0.787	2.576	<0.001	0.889	1.291	0.0028
	Trt-2	0.881	0.943	7.85	0.913	0.867	0.74	0.969	0.415	1.00
	Trt-3	0.880	0.943	7.69	0.914	0.833	0.80	0.976	0.389	1.00
	Trt-4	0.817	0.869	12.92	0.843	2.351	<0.001	0.916	1.068	0.24
	Trt-5	0.811	0.936	12.15	0.838	2.078	<0.001	0.942	0.902	0.86
	Trt-6	0.772	0.810	12.25	0.806	2.114	<0.001	0.906	1.008	0.46
	Trt-All	0.880	0.954	6.45	0.916	0.586	0.99	0.981	0.279	1.00
WSF2	Trt-1	0.826	0.939	8.90	0.874	1.117	0.31	0.961	0.503	1.00
	Trt-2	0.807	0.959	9.27	0.855	1.211	0.20	0.960	0.544	1.00
	Trt-3	0.824	0.936	8.80	0.867	1.091	0.35	0.960	0.505	1.00
	Trt-4	0.802	0.893	9.37	0.843	1.236	0.17	0.949	0.589	1.00
	Trt-5	0.731	0.849	12.87	0.768	2.334	<0.001	0.878	1.204	0.021
	Trt-6	0.602	0.762	17.76	0.645	4.446	<0.001	0.755	2.409	<0.001
	Trt-All	0.854	0.938	7.53	0.894	0.802	0.84	0.970	0.377	1.00
WSF3	Trt-1	0.835	0.907	8.60	0.875	1.041	0.43	0.963	0.497	1.00
	Trt-2	0.814	0.936	9.57	0.850	1.289	0.13	0.962	0.576	1.00
	Trt-3	0.846	0.948	8.26	0.884	0.961	0.58	0.971	0.435	1.00
	Trt-4	0.865	0.946	7.10	0.905	0.709	0.94	0.975	0.331	1.00
	Trt-5	0.754	0.862	12.24	0.791	2.111	<0.001	0.894	1.085	0.19
	Trt-6	0.685	0.639	17.64	0.720	4.384	<0.001	0.807	2.277	<0.001
	Trt-All	0.867	0.942	6.96	0.907	0.683	0.96	0.978	0.316	1.00

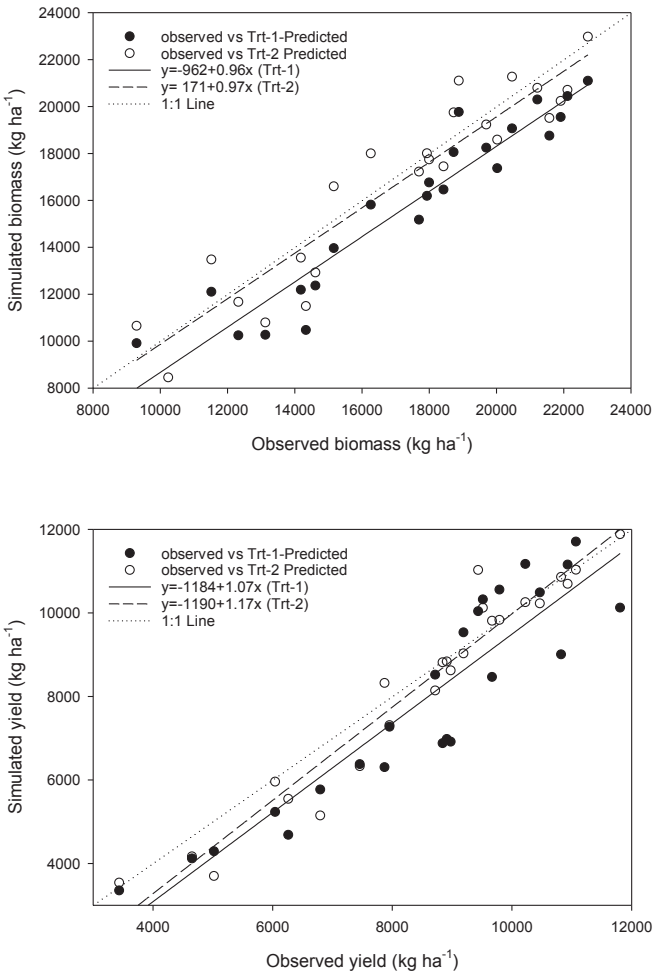

Fig. 1. Regression analysis of predicted biomass and yield with optimized cultivar parameters from Trt-1 and Trt-2 for WSF1 (see Tables 4 and 5).

3. Results and discussion

3.1. Optimized crop cultivar parameters

Calibrated crop cultivar parameters varied for each subset of data used for optimization (Table 2). Among the six treatments (across four years), P1 (thermal time from seedling emergence to end of juvenile phase) had the least CV, followed by PHINT (thermal time required for a leaf tip to emerge) and P5 (thermal time from silking to physiological maturity). P2 (Day length sensitivity coefficient) had the highest CV. G2 (Potential kernel number per plant) and G3 (Potential kernel growth rate) were closely related in yield formation. As a result, G2 and G3 also had high CV among the six irrigation treatments (Table 2). In addition, we noticed smaller CV for WSF3 and highest CV for WSF1. When fitting all the treatments in one year, similar trends in CV among years were observed with lowest CVs for P1, PHINT, and P5, which defines crop phenology. However, there were no obvious trends in CV of fitted parameters among the three water stress functions (Table 3). Except for P1 and P2, all the other parameters had reached the upper boundary for some optimization scenarios, which suggests that the fitted parameters may not be reliable in these cases. Except for P2, the CV among treatments (Trt-1 to Trt-6, Table 2) was smaller than that among years (Year-2008 to Year-2011, Table 3), which suggests it may be better to fit a treatment across years than to fit all treatments in one year as far as parameter stability concerns. Regardless of water stress functions used, fitting all the datasets derived the most reliable model parameters, which suggests that it is best to use all data from one study for model calibration (Ma et al., 2012b).

3.2. Statistics of prediction after calibration using one treatment across four years

After calibrating with one treatment across four years, biomass was generally better simulated than yield based on statistics given in Tables 4 and 5. All r^2s for prediction were greater than 0.83 for all treatments and all water stress functions (Tables 4 and 5). A majority of d-index values were greater than 0.7, which would be 'satisfactory' according to Saseendran et al. (2010) and Ma et al. (2011). RRMSEs range from 7 to 11% for biomass prediction and 7 to 18% for yield prediction. Although these statistics are all acceptable (Ma et al., 2011), they cannot be used to statistically discriminate among the optimization options. In addition, they did not consider experimental uncertainty nor simulation uncertainty. By considering experimental uncertainty and using the d1-index from Harmel and Smith (2007), we found an increase in the index values for all the treatments and water stress functions but could not statistically determine which treatment or which water stress functions provided the best biomass and yield prediction. The d1-index showed the exact same trends for predicted biomass and yield as the d-index. To investigate the effects of simulation uncertainty on model performance, we used CV of 0.075 as simulation uncertainty in this study, which is close to the average CV of experimental uncertainty of yield and biomass. This CV for simulation uncertainty was also reasonable according to Ma et al. (2016) who found that the maximum simulation uncertainty had a CV of 0.07 for both yield and biomass due to spatial variability in soil field capacity (model input data). In addition, if we lumped all the simulation results together in this study from both water stress functions (model structure uncertainty) and calibration datasets (parameterization uncertainty), we obtained a CV of 0.065 for biomass and a CV of 0.107 for yield. Given all the three uncertainties in model simulation, using a CV of 0.075 for our study was reasonable. As a result, if we considered both experimental uncertainty and simulation uncertainty having the same CV of 0.075, the d2-index of Harmel et al. (2010) provided even high index values but with the same trends in model performance among treatments and water stress functions as d-index and d1-index, except for no significant differences among water stress functions for biomass prediction. Thus, no statistical power was added to differentiate the various optimization options.

Therefore, we evaluated the F-test from Ma et al. (2012a) (F1 in Tables 4 and 5). For WSF1, the F-test showed significant differences between experimental and simulated biomass for Trt-1 and Trt-4 when only experimental uncertainties were considered (F1). However, these differences became insignificant when both experimental and simulation uncertainties were taken into account (F2 in Table 4). For yield prediction, only Trt-2 and Trt-3 provided good prediction at significance level of 0.05 when experimental uncertainty was considered. After considering simulation uncertainty, Trt-1 still did not provide significantly good prediction of yield ($p < 0.01$, Table 5). Thus, using only full irrigation treatments for model calibration may not always be the best strategy. Comparisons should be made between water-deficit treatments and well-watered treatments (Boote, 1999) to improve model responses to water deficits. For WSF2, only Trt-1 and Trt-5 showed significantly satisfactory prediction of biomass with only experimental uncertainties considered ($p > 0.05$, Table 4), but all treatments provided good prediction of biomass after considering both

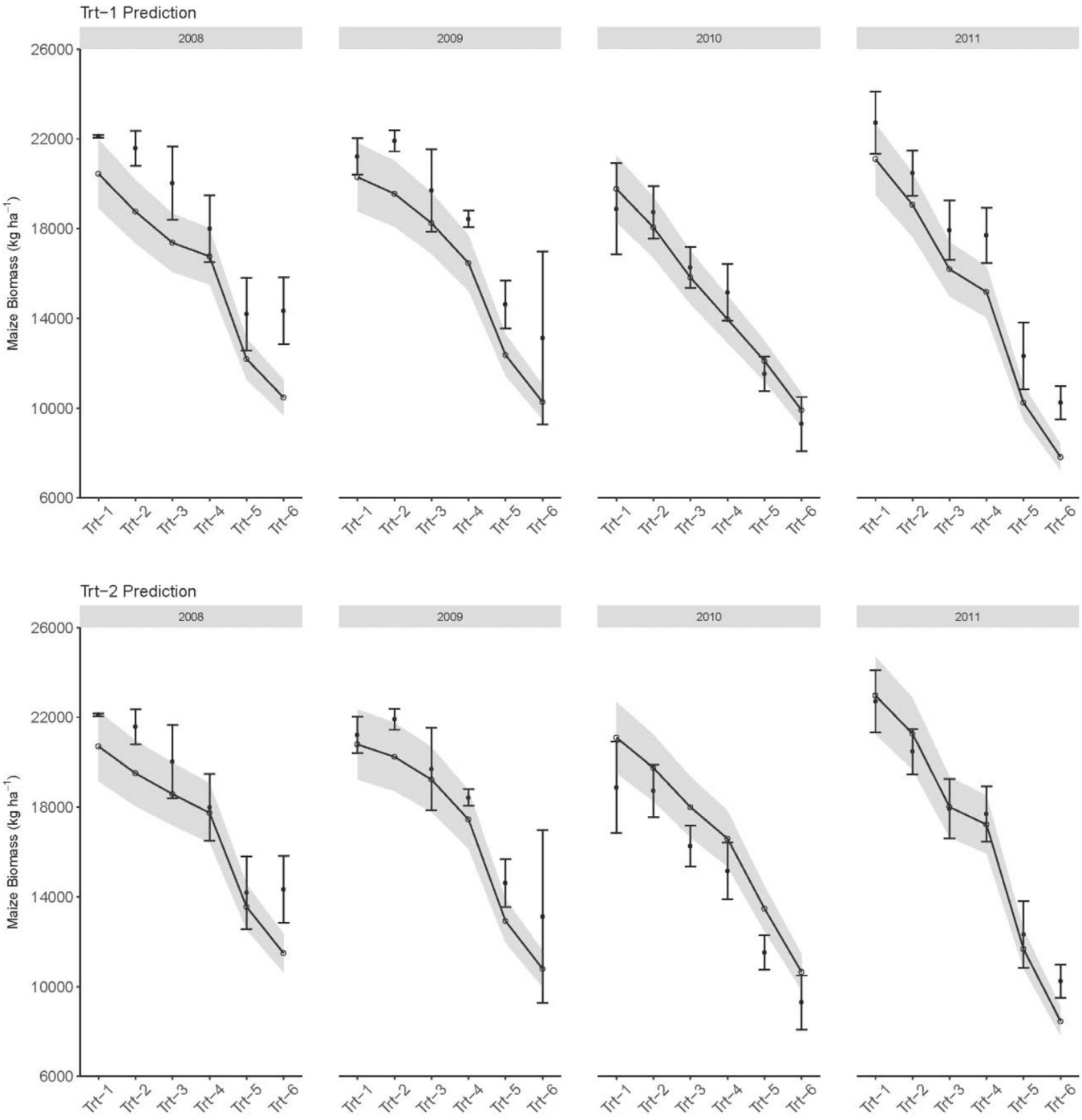

Fig. 2. Experimental and simulated biomass from Trt-1 and Trt-2 optimized cultivars for WSF1. Standard deviation (SD) of simulation was calculated based on a CV (coefficient of variation) of 0.075, and the average CV of experimental uncertainty (See Table 4). The vertical bars and shaded areas are ±1 SD around the mean for experimental and simulation results, respectively.

experimental and simulation uncertainties ($p > 0.05$). For yield, regardless of experimental or simulation uncertainties, Trt-5 and Trt-6 did not provide statistically satisfactory yield prediction ($p < 0.05$, Table 5). For WSF3, Trt-1, Trt-2, and Trt-6 did not provide good biomass prediction with only experimental uncertainty ($p < 0.05$), but showed good prediction when both experimental and simulation uncertainties were considered ($p > 0.5$, Table 4). Yield prediction showed the same trends as in the cases of WSF1, except for satisfactory prediction for Trt-5 after considering simulation uncertainty.

Thus, the modified F-tests were able to discern the goodness of model prediction among different optimization options and stress functions. Based on RRMSE and d-index, Saseendran et al. (2014) concluded that WSF2 and WSF3 were superior to WSF1 based on simulated biomass, yield, and leaf area index after manually calibrating for Trt-1 using WSF1. In our study, we found that if both experimental and simulation uncertainties were taken into account, there was no significant difference among the three water

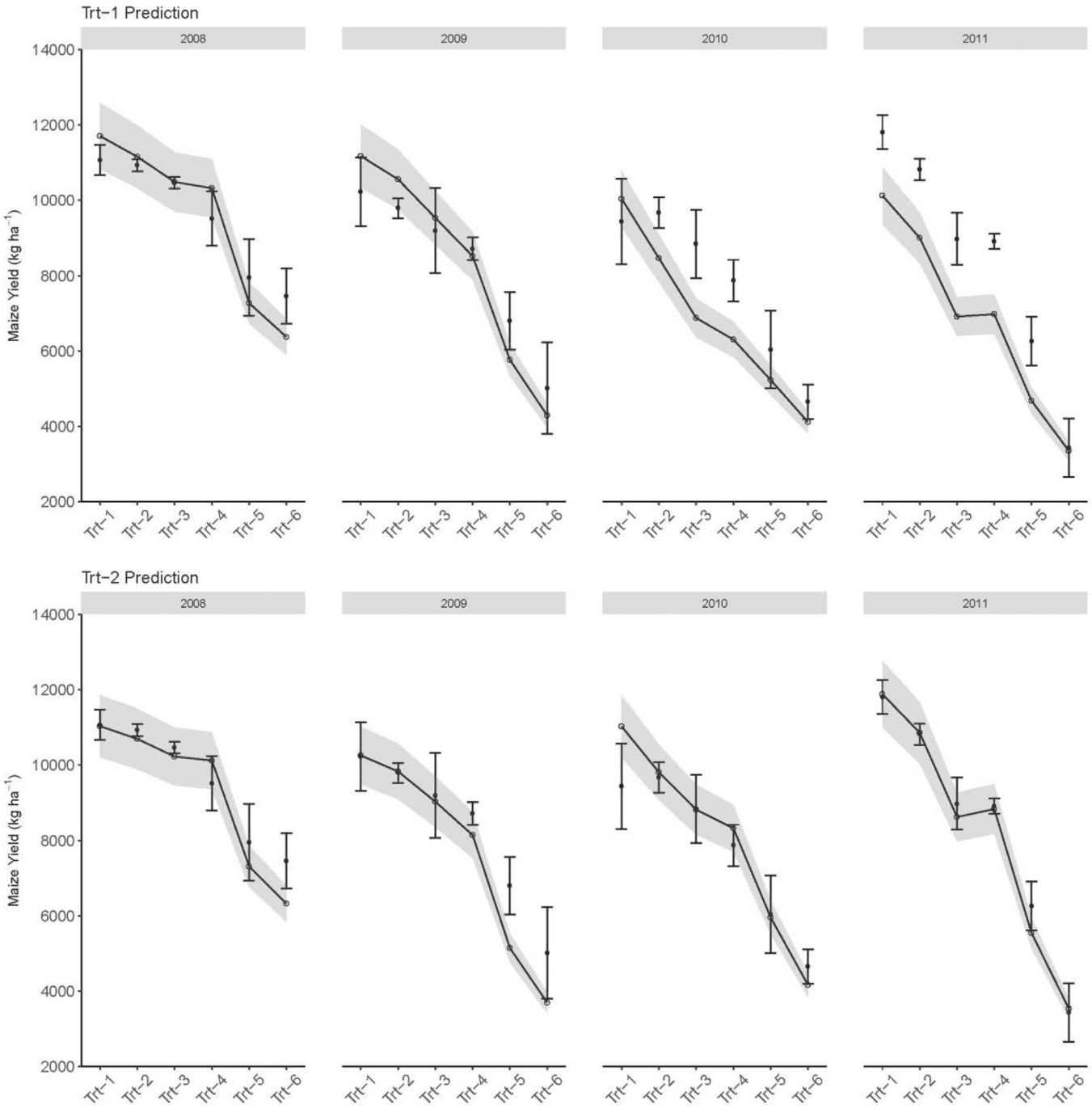

Fig. 3. Experimental and simulated yield from Trt-1 and Trt-2 optimized cultivars for WSF1. Standard deviation (SD) of simulation was calculated based on a CV (coefficient of variation) of 0.075, and the average CV of experimental uncertainty (See Table 5). The vertical bars and shaded areas are ±1 SD around the mean for experimental and simulation results, respectively.

stress functions. Among the treatments used for model calibration, Trt-2 and Trt-3 of WSF1, Trt-1 of WSF2, and Trt-3 and Trt-4 of WSF3 provided statistically good prediction for both biomass and yield when only experimental uncertainties were considered based on the F-tests ($p > 0.1$, Tables 4 and 5). When both experimental and simulation uncertainties were considered, only Trt-1 of WSF1, Trt-5 and Trt-6 of WSF2, and Trt-6 of WSF3 did not satisfy goodness of prediction for yield ($p < 0.05$, Tables 4 and 5). When all the treatments were used for model calibration, there was no difference in goodness of yield and biomass prediction among the three water stress functions statistically ($p > 0.305$) (Tables 4 and 5).

To show that the F-test was better than the traditional statistics, we plotted the simulated and observed biomass predicted with cultivar parameters from Trt-1 and Trt-2 for WSF1 (Fig. 1). As shown in Table 4, Trt-1 predicted biomass had a higher r^2 than Trt-2 predicted biomass, although all other traditional statistics (d-index and RRMSE) were better for Trt-2 prediction. Thus, multiple traditional statistics should be used to evaluate model performance (Moriasi et al., 2007). However, the F-test not only showed better performance of Trt-2 than Trt-1 but also provided a significance

Table 6
F-test, d-index, RRMSE (%), and r^2 for predicting biomass for the three water stress functions with simulation CV of 0.075 for all treatments in one year. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation. For each year, L = 6, N_i = 4.

WSF	Calibration Year	d-index	r^2	RRMSE	Experimental Uncertainty only			Experimental and Simulation Uncertainty		
					d1-index	F1-value	p-value	d2-index	F2-value	p-value
WSF1	Year-2008	0.773	0.828	11.90	0.801	2.005	0.001	0.941	0.888	0.897
	Year-2009	0.632	0.841	17.33	0.656	4.250	<0.001	0.788	2.335	<0.001
	Year-2010	0.690	0.881	13.82	0.721	2.704	<0.001	0.860	1.438	<0.001
	Year-2011	0.823	0.889	9.37	0.857	1.242	0.167	0.960	0.558	1.000
WSF2	Year-2008	0.783	0.914	9.04	0.824	1.156	0.261	0.937	0.548	1.000
	Year-2009	0.681	0.838	12.05	0.889	2.055	<0.001	0.889	1.037	0.346
	Year-2010	0.732	0.836	11.88	0.773	1.999	0.001	0.886	1.028	0.379
	Year-2011	0.784	0.872	8.88	0.831	1.117	0.313	0.940	0.546	1.000
WSF3	Year-2008	0.731	0.821	11.14	0.771	1.757	0.006	0.918	0.808	0.988
	Year-2009	0.741	0.881	10.19	0.789	1.470	0.044	0.922	0.733	0.999
	Year-2010	0.632	0.828	15.32	0.667	3.323	<0.001	0.821	1.792	<0.001
	Year-2011	0.757	0.824	10.41	0.800	1.533	0.029	0.910	0.733	0.999

Table 7
F-test, d-index, RRMSE (%), and r^2 for predicting yield for the three water stress functions with simulation CV of 0.075 for all treatments in one year. WSF1, WSF2, and WSF3 are water stress functions used in RZWQM. CV is coefficient of variation. For each year, L = 6, Ni = 4.

WSF	Calibration Year	d-index	r^2	RRMSE	Experimental Uncertainty only			Experimental and Simulation Uncertainty		
					d1-index	F1-value	p-value	d2-index	F2-value	p-value
WSF1	Year-2008	0.727	0.742	17.31	0.755	4.224	<0.001	0.827	2.010	<0.001
	Year-2009	0.475	0.564	32.97	0.485	15.312	<0.001	0.511	9.437	<0.001
	Year-2010	0.780	0.872	12.75	0.811	2.292	<0.001	0.872	1.578	0.056
	Year-2011	0.822	0.937	9.71	0.858	1.327	0.104	0.968	0.586	1.000
WSF2	Year-2008	0.721	0.754	14.79	0.761	3.083	<0.001	0.845	1.590	<0.001
	Year-2009	0.509	0.510	27.07	0.532	10.326	<0.001	0.603	6.076	<0.001
	Year-2010	0.873	0.936	6.44	0.922	0.585	0.993	0.981	0.279	1.000
	Year-2011	0.837	0.919	8.06	0.880	0.916	0.659	0.972	0.418	1.000
WSF3	Year-2008	0.789	0.888	9.48	0.835	1.267	0.146	0.937	0.611	1.000
	Year-2009	0.653	0.647	214	0.680	5.715	<0.001	0.755	3.114	<0.001
	Year-2010	0.785	0.818	11.57	0.828	1.885	0.003	0.901	0.925	0.796
	Year-2011	0.721	0.857	13.84	0.752	2.670	<0.001	0.917	1.166	0.049

Table 8
Optimized crop parameters fitted for two treatments across four years. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation.

Parameter Name and (Ranges, initial values) for PEST optimization	WSF	CV	Parameter values fitted from two treatments across 4 years				
			Trt-1+Trt-2	Trt-1+Trt-3	Trt-1+Trt-4	Trt-1+Trt-5	Trt-1+Trt-6
P1 - Degree days (base temperature of 8 °C) from seedling emergence to end of juvenile phase (thermal degree days) (100–450, 250).	WSF1	0.027	258.1	258.3	259.0	256.7	242.9
	WSF2	0.046	249.5	251.6	253.1	228.3	234.3
	WSF3	0.048	245.4	250.7	226.6	225.8	244.4
P2 - Day length sensitivity coefficient [the extent (days) that development is delayed for each hour increase in photoperiod above the longest photoperiod (12.5 h) at which development proceeds at maximum rate] (0–1, 0.2).	WSF1	0.635	0.236	0.377	0.176	0.017	0.194
	WSF2	0.334	0.075	0.203	0.181	0.226	0.185
	WSF3	0.635	0.182	0.189	0.022	0.142	0.057
P5 - Degree days (base temperature of 8 °C) from silking to physiological maturity (thermal degree days) (500–1000, 600)	WSF1	0.085	721.2	717.8	587.5	729.7	683.0
	WSF2	0.055	707.3	698.0	700.2	794.2	730.6
	WSF3	0.107	692.6	689.4	860.2	790.7	679.4
G2 - Potential kernel number per plant (440–1000, 900)	WSF1	0.270	985.0	974.0	863.5	440.0	843.7
	WSF2	0.099	756.0	977.9	939.2	971.6	911.6
	WSF3	0.252	966.3	1000.0	793.8	942.0	486.1
G3 - Potential kernel growth rate (mg/(kernel d) (5–16, 6)	WSF1	0.300	6.31	6.43	7.59	12.26	7.88
	WSF2	0.147	7.93	6.33	6.49	5.31	6.15
	WSF3	0.366	6.56	6.36	5.93	5.60	12.03
PHINT - Degree days required for a leaf tip to emerge (thermal degree days) (38–55, 50)	WSF1	0.080	45.7	43.8	50.4	46.3	53.2
	WSF2	0.020	49.2	48.3	48.7	50.9	49.6
	WSF3	0.043	48.1	50.1	47.0	52.2	51.5

Table 9
F-test, d-index, RRMSE (%), and r^2 for predicting biomass for the three water stress functions with simulation CV of 0.075 for two treatments across four years. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation. For each calibration, L = 8, N_i = 4.

WSF	Calibration Treatment	d-index	r^2	RRMSE	Experimental Uncertainty only			Experimental and Simulation Uncertainty		
					d1-index	F1-value	p-value	d2-index	F2-value	p-value
WSF1	Trt-1+Trt-2	0.816	0.911	0.085	0.862	1.025	0.459	0.950	0.508	1.000
	Trt-1+Trt-3	0.835	0.898	0.075	0.877	0.796	0.853	0.968	0.381	1.000
	Trt-1+Trt-4	0.791	0.889	0.093	0.839	1.236	0.173	0.939	0.616	1.000
	Trt-1+Trt-5	0.799	0.908	0.089	0.846	1.136	0.286	0.944	0.569	1.000
	Trt-1+Trt-6	0.745	0.911	0.118	0.786	1.986	0.001	0.894	1.034	0.357
WSF2	Trt-1+Trt-2	0.768	0.889	0.094	0.809	1.254	0.156	0.940	0.586	1.000
	Trt-1+Trt-3	0.757	0.876	0.101	0.795	1.441	0.052	0.929	0.668	1.000
	Trt-1+Trt-4	0.759	0.879	0.099	0.798	1.402	0.066	0.930	0.651	1.000
	Trt-1+Trt-5	0.782	0.860	0.089	0.826	1.117	0.313	0.947	0.538	1.000
	Trt-1+Trt-6	0.799	0.919	0.080	0.845	0.914	0.662	0.954	0.439	1.000
WSF3	Trt-1+Trt-2	0.794	0.888	0.085	0.838	1.022	0.464	0.954	0.482	1.000
	Trt-1+Trt-3	0.774	0.859	0.095	0.819	1.269	0.145	0.936	0.593	1.000
	Trt-1+Trt-4	0.801	0.865	0.088	0.847	1.108	0.325	0.939	0.544	1.000
	Trt-1+Trt-5	0.809	0.890	0.085	0.854	1.022	0.465	0.942	0.505	1.000
	Trt-1+Trt-6	0.790	0.884	0.086	0.838	1.058	0.402	0.946	0.501	1.000

level test (p = 0.47 for Trt-2 and p = 0.003 for Trt-1). As shown in Fig. 1, Trt-2 predictions were closer to the 1:1 line than Trt-1 predictions. In addition, the liner regression of Trt-1 data points failed the constant variance test (p = 0.016), which suggested a non-zero residual for the linear regression. For yield, Trt-2 predicted higher r^2 than Trt-1, which was in agreement with the F-test (Fig. 1; Table 5). Fig. 2 showed the same biomass data but plotted for treatment by year to visualize the differences between Trt-1 and Trt-2 predicted biomass. Trt-1 under-predicted biomass for 2008, 2009, and 2011 considerably, whereas Trt-2 under predicted biomass for 2008 but over predicted biomass in 2010. The under prediction of biomass by Trt-1 was so severe that both experimental and simulation uncertainties had to be considered to make a statement that simulated and observed results were not significantly different. For yield, Trt-1 under-predicted yield in 2010 and 2011 compared to Trt-2 (Fig. 3) and the simulation results were significantly different from observed yield even if both experimental and simulation uncertainties were considered (p = 0.0028, Table 5).

3.3. Statistics of prediction after calibration using all treatments in one year

Ma et al. (2012b) suggested calibrating a model with all treatments in one year so that the treatment effects can be taken into account in model optimization. However, in this study we found the opposite. As shown in Tables 6 and 7, all statistics were inferior to those obtained when calibrating the model with one treatment from four years. Similar to treatment options in model calibration, it was not possible to discriminate among the calibration options (Year-2008, Year-2009, Year-2010, and Year 2011) using the d-index, r^2, or RRMSE, although a combination of these statistics might be helpful (Ma et al., 2011; Moriasi et al., 2007). However, we did see Year-2011 provided statistically good prediction of both biomass and yield for WSF1 and WSF2 based on the F-tests when experimental uncertainties were considered. When both experimental and simulation uncertainties were taken into account, Year-2011 of WSF1, Year-2010 and Year-2011 of WSF2, and Year-2008 and Year-2011 of WSF3 provided statistically satisfactory goodness of model predictions (Tables 6 and 7).

Looking at the yield and biomass variations among the years, we found that Year-2011 had the most differences between highest and lowest treatments for both biomass (10235–22721 kg ha^{-1}) and yield (3434–11809 kg ha^{-1}). This may be the reason that calibrated parameters from Year-2011 provided the best goodness of prediction because the calibrated cultivar parameters had compensated for the water stress endured by plants. Therefore, when multiple

Table 10
F-test, d-index, RRMSE (%), and r^2 for predicting yield for the three water stress functions with simulation CV of 0.075 for two treatments across four years. WSF1, WSF2, and WSF3 are water stress functions (WSF) used in RZWQM. CV is coefficient of variation. For each calibration, L = 8, N_i = 4.

WSF	Calibration Treatment	d-index	r^2	RRMSE	Experimental Uncertainty only			Experimental and Simulation Uncertainty		
					d1-index	F1-value	p-value	d2-index	F2-value	p-value
WSF1	Trt-1+Trt-2	0.834	0.938	0.089	0.870	1.121	0.307	0.966	0.547	1.000
	Trt-1+Trt-3	0.878	0.951	0.068	0.920	0.649	0.975	0.975	0.308	1.000
	Trt-1+Trt-4	0.865	0.945	0.077	0.903	0.845	0.779	0.967	0.400	1.000
	Trt-1+Trt-5	0.816	0.927	0.092	0.855	1.191	0.218	0.953	0.589	1.000
	Trt-1+Trt-6	0.812	0.931	0.101	0.848	1.438	0.053	0.946	0.708	1.000
WSF2	Trt-1+Trt-2	0.834	0.939	0.087	0.877	1.068	0.386	0.964	0.486	1.000
	Trt-1+Trt-3	0.826	0.956	0.085	0.874	1.014	0.479	0.965	0.464	1.000
	Trt-1+Trt-4	0.832	0.949	0.084	0.879	0.982	0.536	0.965	0.453	1.000
	Trt-1+Trt-5	0.854	0.913	0.078	0.900	0.855	0.764	0.970	0.400	1.000
	Trt-1+Trt-6	0.851	0.922	0.077	0.894	0.847	0.777	0.966	0.402	1.000
WSF3	Trt-1+Trt-2	0.846	0.935	0.075	0.892	0.800	0.846	0.974	0.367	1.000
	Trt-1+Trt-3	0.852	0.952	0.077	0.893	0.836	0.794	0.973	0.381	1.000
	Trt-1+Trt-4	0.843	0.886	0.087	0.886	1.079	0.369	0.958	0.519	1.000
	Trt-1+Trt-5	0.867	0.921	0.072	0.909	0.720	0.993	0.978	0.344	1.000
	Trt-1+Trt-6	0.816	0.939	0.086	0.861	1.056	0.406	0.970	0.475	1.000

treatments in one year were used for model calibration, selecting the one showing the most treatment effects was warranted.

3.4. Statistics of prediction after calibration with two treatments in all years simultaneously

To evaluate model robustness when two treatments from all four years were used for calibration simultaneously, we selected Trt-1 plus another treatment. As shown in Table 8, the obtained cultivar parameters had a lower CV compared to the previous two optimization options. Except for G2 under WSF3, all the parameters were within their respective specified ranges. Goodness of prediction of these parameters was much improved compared to those obtained from optimizing either one treatment across all years or all treatments in one year (Tables 9 and 10). The improvement seems to be more obvious for yield than for biomass for all statistics. When only experimental uncertainty was considered, there were no significant differences between predicted and observed biomass and yield regardless of water stress functions used ($p > 0.05$), except for biomass prediction under WSF1 with parameters from Trt-1+Trt-6 (Table 9). When both experimental and simulation uncertainties were taken into account, all predictions were acceptable, which suggests the three water stress functions and the five optimization options were equally effective. Therefore, the conclusion on water stress functions (Saseendran et al., 2014) may not be valid when more rigorous statistics, such as the modified F-test, were used. The results also showed that model calibration would be more stable and reliable if two or more treatments were used in model calibration (Fensterseifer et al., 2017) (see Table 10).

4. Conclusion

In this study, a modified F-test was developed to account for experimental and simulation uncertainty. The experimental uncertainty was generally taken from measurement uncertainty in data used for calibration and evaluation. Simulation uncertainties might be from model structure, model inputs, as well as model parameters. An irrigation study was used to exemplify these traditional and enhanced goodness of fit statistics and their application for objectively differentiating goodness of model prediction. Based on the results, goodness of model prediction heavily depends on which sub-datasets were used for model calibration and which outputs were compared (yield or biomass). When taking into account both experimental and simulation uncertainties, biomass was well simulated regardless of treatments and water stress functions used for calibration. In general, biomass was better predicted than yield. Using one treatment across years for model calibration seemed to be superior to calibration using all treatments in one year, especially when both experimental and simulation uncertainties were taken into account. It is recommended to use two or more treatments for model calibration to obtain most reliable model parameters and better prediction.

We found that accounting for uncertainty in both experimental and simulation results increased the d-index for all optimization options and water stress functions. But such an increase could not be used to discern the goodness of model prediction. The power of F-test depends on the coefficient of variance (CV) of both experimental and simulation results. If the CV of simulation uncertainty is smaller than 0.075, the modified F-test will reject more simulation results and show more significant differences between simulated and observed results. In addition, if other model outputs (e.g., leaf area index, soil water content, etc.) are used in the statistics, the conclusion may change. In the paper, we analyzed statistical significance for each yield and biomass separately, but an overall model performance may be developed by pooling all the outputs together in the statistics. In conclusion, due to its ability to form a statistical significance test, the modified F-test should be recommended for more rigorous model evaluation than the traditional simple statistics when experimental uncertainty and/or simulation uncertainty are available. By properly calibrating the model, all the three water stress functions provided similar goodness of prediction for yield and biomass in this study. Where there were more treatment-year data available, using more data for calibration would increase model performance and predictability.

Acknowledgement

The authors wish to thank Dr. Thomas Trout, retired USDA-ARS employee, for sharing the data at https://data.nal.usda.gov/dataset/usda-ars-colorado-maize-water-productivity-dataset-2008-2011.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.envsoft.2018.03.011.

References

Ahuja, L.R., Rojas, K.W., Hanson, J.D., Shaffer, M.J., Ma, L., 2000. Root Zone Water Quality Model. Water Resources Publications LLC, Highlands Ranch, CO, p. 358.
Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop Evapotranspiration: Guidelines for Computing Crop Water Requirement, vol. 56. United Nations Food and Agriculture Organiation, Irrigation and Drainage paper, Rome, Italy, p. 300.
Allen, R.G., Walter, I.A., Elliott, R.L., Howell, T.A., Itenfisu, D., Jensen, M.E., Snyder, R.L., 2005. The ASCE Standardized Reference Evapotranspiration Equation. American Society of Civil Engineers, Reston, VA.
Boote, K.J., 1999. Concepts for calibrating crop growth models. In: Hoogenboom, G., Wilkens, P.W., Tsuji, G.Y. (Eds.), DSSAT Version 3, vol. 4. University of Hawaii, Honolulu, pp. 179–199.
Doherty, J., 2010. PEST: Model-independent Parameter Estimation. Watermark Numerical Computing, Australia. http://www.pesthomepage.org.
Fensterseifer, C.A., Streck, N.A., Baigorria, G.A., Timilsina, A.P., Zanon, A.J., Cera, J.C., Rocha, T.S.M., 2017. On the number of experiments required to calibrate a cultivar in a crop model: the case of CROPGRO-soybean. Field Crop. Res. 204, 146–152.
Harmel, R.D., Smith, P.K., 2007. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. J. Hydrol. 337 (3), 326–336.
Harmel, R.D., Smith, P.K., Migliaccio, K.W., 2010. Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. Trans. ASABE (Am. Soc. Agric. Biol. Eng.) 53, 55–63.
Harmel, R.D., Smith, P.K., Migliaccio, K.W., Chaubey, I., Douglas-Mankin, K.R., Benham, B., Shukla, S., Muñoz-Carpena, R., Robson, B.J., 2014. Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: a review and recommendations. Environ. Model. Software 57, 40–51.
Kersebaum, K.C., Wurbs, A., De Jong, R., Campbell, C.A., Yang, J., Zentner, R.P., 2008. Long-term simulation of soilcrop interactions in semiarid southwestern Saskatchewan, Canada. Eur. J. Agron. 29 (1), 1–12.
Krause, P., Boyle, D.P., Base, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.
Ma, L., Ahuja, L.R., Saseendran, S.A., Malone, R.W., Green, T.R., Nolan, B.T., Bartling, P.N.S., Flerchinger, G.N., Boote, K.J., Hoogenboom, G., 2011. A Protocol for parameterization and calibration of RZWQM2 in field research. In: Ahuja, L.R., Ma, L. (Eds.), Methods of Introducing System Models into Agricultural Research, Advances in Agricultural Systems Modeling, vol. 2. Soil Science Society of America, Madison, WI, pp. 1–64.
Ma, L., Ahuja, L.R., Trout, T.J., Nolan, B.T., Malone, R.W., 2016. Simulating maize yield and biomass with spatial variability of soil field capacity. Agron. J. 108, 171–184.
Ma, L., Ahuja, L.R., Nolan, B.T., Malone, R.W., Trout, T.J., Qi, Z., 2012a. Root Zone water quality model (RZWQM 2): model use, calibration, and validation. Trans. ASABE (Am. Soc. Agric. Biol. Eng.) 55 (4), 1425–1446.
Ma, L., Hoogenboom, G., Ahuja, L.R., Ascough II, J.C., Anapalli, S.S., 2006. Evaluation of RZWQM-CERES-maize hybrid model for maize production. Agric. Syst. 87, 274–295.
Ma, L., Trout, T.J., Ahuja, L.R., Bausch, W., Saseendran, S.A., Malone, R.W., Nielsen, D.C., 2012b. Calibrating RZWQM2 model for maize responses to deficit irrigation. Agric. Water Manag. 103, 140–149.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE (Am. Soc. Agric. Biol. Eng.) 50 (3),

885–900.
Nimah, M.N., Hanks, R.J., 1973. Model for estimating soil water, plant and atmospheric inter relations: I. description and sensitivity. Proc. Soil Sci. Soc. Am. 37, 522–527.
Ritchie, J.T., 1998. Soil water balance and plant water stress, pp. 41–54. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), Understanding Options for Agricultural Production. Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 41–54.
Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol 480, 33–45.
Saseendran, S.A., Nielsen, D.C., Ma, L., Ahuja, L.R., Vigil, M.F., 2010. Simulating alternative dryland rotational cropping systems in the central Great Plains with RZWQM2. Agron. J. 102, 1521–1534.
Saseendran, S.A., Ahuja, L.R., Ma, L., Nielsen, D.C., Trout, T.J., Andales, A.A., Chávez, J.L., Ham, J., 2014. Enhancing the water stress factors for simulation of corn (*Zea mays* L.) in RZWQM2. Agron. J. 106, 81–94.
Trout, T.J., Bausch, W.C., 2017. USDA-ARS Colorado maize water productivity data set. Irrigat. Sci. 35, 241–249. https://doi.org/10.1007/s00271-017-0537-9.
Wackerly, D.D., Mendenhall III, W., Scheaffer, R.L., 2008. Mathematical Statistics with Applications, seventh ed. Brooks/Cole Cengage Learning, Belmont, Cal.
Whitmore, A.P., 1991. A method for assessing the goodness of computer simulation of soil processes. J. Soil Sci. 42, 289–299.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.
Yen, H., Wang, X., Fontane, D.G., Harmel, R.D., Arabi, M., 2014. A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modeling. Environ. Model. Software 54, 211–221.
Yen, H., Hoque, Y.M., Wang, X., Harmel, R.D., 2016. Applications of explicitly incorporated/post-processing measurement uncertainty in watershed modeling. J. Am. Water Resour. Assoc. 52, 523–540.