Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences
Open Access
Discussions

Bayesian performance evaluation of evapotranspiration models for an arid region in northwestern China

Guoxiao Wei [1, 2,*], Xiaoying Zhang [3,4,*], Ming Ye [5], Ning Yue [1,2], Fei Kan [1,2]

[1] Key Laboratory of Western China's Environmental System (Ministry of Education), Lanzhou University, China, 730000

[2] School of Earth and Environmental Sciences, Lanzhou University, China, 730000

[3] Institute of Groundwater and Earth Sciences, Jinan University, China, 510632

[4] Construct Engineering College, Jilin University, China, 130400

[5] Department of Earth, Ocean, and Atmospheric Science, Florida State University, USA, 32306

* Corresponding author: xiaoyingzh@jlu.edu.cn; gxwei@lzu.edu.cn.

Abstract

Evapotranspiration (ET) is a major component of the land surface process involved in energy fluxes and balance, especially in the hydrological cycle of agricultural ecosystems. While many models have been developed to estimate ET, there has been no agreement on which model has the best performance. In this study, we evaluate four widely used ET models (i.e., the Shuttleworth Wallace (SW) model, Penman-Monteith (PM) model, Priestley-Taylor and Flint-Childs (PT-FC) model, and Advection-Aridity (AA) model) by using half-hourly ET observations obtained at a spring maize field in an arid region. The model evaluation is based on Bayesian model comparison and ranking using the Bayesian model evidence (BME), which balances between goodness-of-fit to data and model complexity. The BME-based model ranking (from the best to the worst) is SW, PM, PT-FC, and AA. The residuals between observations and corresponding model simulations are also analyzed, and the same model ranking is also obtained by using residual-based statistics, i.e., the coefficient of determination (R^2), index of agreement (IA), root mean square error (RMSE) and model efficiency (EF). The PM and SW models overestimate ET, whereas the PT-FC and AA models underestimate ET in the study period. The four models also underestimate ET during the periods of partial crop cover. Especially during the late maturity stage, the PT-FC and AA models consistently produce an underestimation, and provide the worst simulated ET. As a result, at the half-hourly time scale, the SW model is the best model and recommend as the first choice for evaluating ET of spring maize in arid desert oasis areas.

Keywords: Evapotranspiration; Bayesian analysis; Penman–Monteith; Shuttleworth-Wallace; Maize

1. Introduction

Surface energy fluxes are an important component of Earth's global energy budget and a primary

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

determinant of surface climate. Evapotranspiration (ET), as a major energy flux process for energy balance, accounts for about 60 ~ 65% of the average precipitation over the surface of the Earth. In agricultural ecosystems, more than 90% of the total water losses are due to ET (Brutsaert, 2005). Therefore, ET estimation is crucial to a wide range of problems in hydrology (Xu and Singh, 1998), ecology, and global climate change (Morison et al., 2008). In practice, much of our understanding of how land surface processes and vegetation affect weather and climate is based on numerical modeling of surface energy fluxes and the atmospherically-coupled hydrological cycle (Bonan, 2008). Several models are commonly used in agricultural systems to evaluate ET. The Penman-Monteith (PM) and Shuttleworth-Wallace (SW) models are physically sound and rigorous (Zhu et al., 2013), and thus widely used to estimate ET for seasonally varied vegetations. The models consider the relationships among net radiation, all kinds of heat flux (such as latent heat sensible heat, and heat from soil and canopy), and surface temperature. The Priestley-Taylor and Flint-Childs (PT-FC) model (based on radiation) and the advection-aridity (AA) model (based on meteorological variables) have also been widely used because they require a small amount of ground-based measurements for setting up the models (Ershadi et al., 2014). Evaluating the performance of these four models is the focus of this study.

These ET models are generally complex for the coupling of the land surface and atmospheric processes, and high-dimensional with a large number of parameters. Modelers are challenged by how to compare the competing models and how to evaluate the mismatch between model simulations and corresponding observed surface-atmosphere water flux (Legates, 1999). Moreover, how to choose a criterion to reliably evaluate model performance is another crucial issue. Both non-Bayesian analysis (Szilagyi and Jozsa, 2008; Vinukollu et al., 2011; Li et al., 2013; Ershadi et al., 2015) and Bayesian analysis have been used for evaluating model performance (Zhu et al., 2014; Chen et al., 2015; Liu et al., 2016; Zhang et al., 2017; Elshall et al., 2018; Samani et al., 2018; Zeng et al., 2018) . These quantitative criteria used for model evaluation and selection include residual-based measures (e.g., regression line slope and mean bias error, MBE), squared residual-based measures (e.g., coefficient of determination, R^2), root mean square error (RMSE), model efficiency (EF), and index of agreement (IA). Li et al. (2013) compared the maize ET estimates given by PM, SW and adjusted SW models under film-mulching conditions in an arid region of China. They found that the half-hourly ET was overestimated by 17% by the SW model, with relatively high MBE, RMSE, and lower R^2 and IA. In contrast, the PM and MSW models underestimated the daily ET by 6% and 2%, respectively, during the entire experimental period of 116 days. Therefore, the performance of PM and MSW models are better than that of the SW model in their case. Ershadi et al. (2014) evaluated the surface energy balance system (SEBS), PM, PT-JPL (modified Priestley–Taylor model, similar to the PT-FC) and AA models. Based on the average value of EF and RMSE, the model rank from the worst to the best was AA, PM, SEBS, and PT-JPL. Ershadi et al. (2015) also evaluated model response to the different formulations of aerodynamic and surface resistances against global FLUXNET data. The results showed significant variability in model performance among and within biome types.

The Bayesian model evidence (BME), also known as marginal likelihood, measures the average fit of a model to the data over a model's parameter space. When comparing several alternative conceptual models, the model with the largest marginal likelihood is selected as the best model (Lartillot and Philippe, 2006). BME can thus be used for evaluating the model fit (over the parameter space) and for comparing alternative models. In previous studies, Bayesian information criterion (BIC; Kashyap, 1982) or Kashyap

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

information criterion (KIC; Schwarz 1978) were used to approximate BME for reducing computational cost of evaluating BME (Ye et al., 2004). However, these approximations have theoretical and computational limitations (Ye et al., 2008; Xie, 2011; Schöniger et al., 2014), and a numerical evaluation (not an approximation) of BME is necessary, especially for complex models (Lartillot and Philippe, 2006). Lartillot and Philippe (2006) advocated the use of thermodynamic integration (TI) for estimating BME, which is also known as path sampling (Gelman and Meng, 1998; Neal, 2000), to avoid sampling solely in the prior or posterior parameter space. TI uses samples that are systematically generated from the prior to the posterior parameter space by conducting path sampling with several discrete power coefficient values (Liu et al., 2016). It is both mathematically rigorous and more accurate than the generally used harmonic mean method (Xie et al., 2011).

While many statistical criteria have beed used to evaluate different ET models, BME has not beed used for evaluating the ET models. It remains to be determined whether BME can be used to compare and select the best model and whether BME can provide an unbiased view of the performance of the models. Furthermore, most Bayesian applications have focused on the calibration of individual models and comparison of alternative models using these statistical measures, with little attention given to the Bayesian model comparison. Model calibration, comparison, and analysis underlying the Bayesian paradigm has been much less used in the evaluation of ET models than in other areas of environmental science.

In this study, the Bayesian approach was used to calibrate and evaluate the four ET models (PM, SW, PT-FC, and AA) based on an experiment over a spring maize field in an arid area of northwest China, from 3 June to 27 September 2014. The ET models were calibrated using the DiffeRential Evolution Adaptive

Metropolis (DREAM) algorithm. The objectives of the study are as follows: (1) to compare the four models and select the best one using BME; (2) to evaluate various general statistics such as correlation-based measures (R^2), relative error measures (IA and EF), and absolute error measures (such as RMSE and MBE) and to determine whether these methods are efficient and reasonable for evaluating the ET models; (3) to analyze model-data mismatch for better understanding model performance. Using BME for evaluating the ET models has not been reported in the literature. We expect that the study will not only boost the development of model parameterization and model selection but also contribute to the improvement of the ET models.

2. Data and methodology

2.1. Description of the study area

The experiment was conducted at Daman Superstation, located in Zhangye, Gansu, northwest China. Daman Oasis is located in the middle Heihe River basin, which is the second largest inland river basin in the arid region of northwest China. The midstream area of the Heihe River basin is characterized by oases with irrigated agriculture, and is a major zone of water consumption for domestic and agricultural uses. The annual average precipitation and temperature are 125 mm and 7.2 °C (1960–2000), respectively. The annual accumulated temperature (>10 °C) is 3,234 °C, and the annual average potential evaporation is about 2,290 mm. The average annual duration of sunshine is 3,106 h with 148 frost-free days. The predominant soil type is silty-clay loam and the depth of the frozen layer is about 143 mm. The study area is a typical irrigated agriculture region, and the major water resources are the snowmelt from the Qilian Mountains. The maize and spring wheat are the principal crops, Maize is generally sown in late April and harvested in mid-September and is planted with a row spacing of 40 cm and a plant spacing of 30 cm. The plant density is about 66,000 plants per hectare.

2.2. Measurements and data processing

Our observation data were collected from the field observation systems of the Heihe Watershed Allied Telemetry Experimental Research (HiWATER) project as described in Li et al (2013). The observation period was from DOY (day of the year) 154 to DOY 270 in 2014. An open-path eddy covariance (EC) system was installed in a maize field, with the sensors at the height of 4.5 m. Maize is the main crop in the study region, which can supply sufficient planting area to set the EC measurements. The EC data was logged at a frequency of 10 H_Z and then processed with an average time interval of 30 min. Sensible and latent heat fluxes were computed by the EC approach of Baldocchi (2003). Flux data measured by EC were controlled by traditional routes, including three-dimensional rotation (Aubinet et al., 2000), WPL (Webb-Penman-Leuning) density fluctuation correction (Webb et al., 1980), frequency response correction (Xu et al., 2014), and spurious data removal caused by rainfall, water condensation, and system failure. About 85% of the energy balance closure was observed in the EC data (Liu et al., 2011).

Standard hydro-meteorological variables, including rainfall, air temperature, wind speed, and wind direction, were continuously measured at the heights of 3, 5, 10, 15, 20, 30 and 40 m above the ground. Soil temperature and moisture were measured at heights of 2, 4, 10, 20, 40, 80, 120 and 160 cm. Photosynthetically active radiation was measured at a height of 12 m. Net radiation, including downward and upward and longwave radiation, was measured by a four-component net radiometer. An infrared thermometer was installed at a height of 12 m. LAI was measured approximately every 10 days during the growing season.

2.3. Model description

2.3.1 Penman-Monteith (PM) model

The PM model can be formulated as following (Monteith, 1965) and most of the parameters are explained in Appendix A:

$$\lambda E = \frac{\varepsilon A + \left(\rho C_{\mathrm{p}} / \gamma\right) D_{\mathrm{a}} g_{\mathrm{a}}}{\varepsilon + 1 + g_{\mathrm{a}} / g_{\mathrm{s}}} \tag{1}$$

where $\varepsilon = \Delta/\gamma$; and A is defined as:

$$A = R_{\mathrm{n}} - G \tag{2}$$

In the present study, g_a is parameterized as suggested by Leuning (2008) and g_s is defined as:

$$g_{\mathrm{s}} = g_{\mathrm{s}}^{\mathrm{c}} \left[\frac{1 + \dfrac{\tau g_{\mathrm{a}}}{(\varepsilon+1) g_{\mathrm{s}}^{\mathrm{c}}} \left[f - \dfrac{(\varepsilon+1)(1-f) g_{\mathrm{s}}^{\mathrm{c}}}{g_{\mathrm{a}}} \right] + \dfrac{g_{\mathrm{a}}}{\varepsilon g_i}}{1 - \tau \left[f - \dfrac{(\varepsilon+1)(1-f) g_{\mathrm{s}}^{\mathrm{c}}}{g_{\mathrm{a}}} \right] + \dfrac{g_{\mathrm{a}}}{\varepsilon g_i}} \right] \tag{3}$$

where 1-τ is the fraction of the total available energy absorbed by the canopy and by the soil, and τ = exp (-$k_A LAI$); g_i is defined as:

$$g_i = \frac{A}{\left(\rho C_{\mathrm{p}} / \gamma\right) D_{\mathrm{a}}} \tag{4}$$

(Monteith, 1965); $g_{\mathrm{s}}^{\mathrm{c}}$ is expressed as:

$$g_{\mathrm{s}}^{\mathrm{c}} = \frac{g_{\max}}{K_{\mathrm{Q}}} In \left[\frac{Q_{\mathrm{h}} + Q_{50}}{Q_{\mathrm{h}} \exp(-K_{\mathrm{Q}} \mathrm{LAI}) + Q_{50}} \right] \left[\frac{1}{1 + D_{\mathrm{a}} / D_{50}} \right] f(\theta) \tag{5}$$

where $f(\theta)$ is the factor considers water stress and is expressed as:

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences Discussions
Open Access

$$f(\theta)=\begin{cases} 1 & \theta>\theta_{\mathrm{a}} \\ \dfrac{\theta-\theta_{\mathrm{b}}}{\theta_{\mathrm{a}}-\theta_{\mathrm{b}}} & \theta_{\mathrm{b}}<\theta<\theta_{\mathrm{a}} \\ 0 & \theta<\theta_{\mathrm{b}} \end{cases} \tag{6}$$

where θ_a was set as θ_a=0.75 θ_b. Aerodynamic conductance is calculated as:

$$g_{\mathrm{a}}=\frac{k^2 u_{\mathrm{m}}}{\ln\left[(z_{\mathrm{m}}-d)/z_{0\mathrm{m}}\right]\ln\left[(z_{\mathrm{m}}-d)/z_{0\mathrm{v}}\right]} \tag{7}$$

where the quantities d, z_{0m} and z_{0v} are calculated using $d = 2h/3$, z_{0m}= 0.123h and z_{0v} = 0.1z_{0m} (Allen 1998).

2.3.2. Shuttleworth-Wallace (SW) model

The SW model comprises a one-dimensional model of plant transpiration and a one-dimensional model of soil evaporation. The two terms are calculated by the following equations:

$$\lambda \mathrm{ET}=\lambda E+\lambda T=C_{\mathrm{s}}\mathrm{ET}_{\mathrm{s}}+C_{\mathrm{c}}\mathrm{ET}_{\mathrm{c}} \tag{8}$$

$$\mathrm{ET}_{\mathrm{s}}=\frac{\Delta A+\left\{\rho C_{\mathrm{p}}(e_{\mathrm{s}}-e_{\mathrm{a}})-\Delta r_{\mathrm{a}}^{\mathrm{s}}\left(A-A_{\mathrm{s}}\right)\right\}/\left(r_{\mathrm{a}}^{\mathrm{a}}+r_{\mathrm{a}}^{\mathrm{s}}\right)}{\Delta+\gamma\left\{1+r_{\mathrm{s}}^{\mathrm{s}}/\left(r_{\mathrm{a}}^{\mathrm{a}}+r_{\mathrm{a}}^{\mathrm{s}}\right)\right\}} \tag{9}$$

$$\mathrm{ET}_{\mathrm{c}}=\frac{\Delta A+\left\{\rho C_{\mathrm{p}}(e_{\mathrm{s}}-e_{\mathrm{a}})-\Delta r_{\mathrm{a}}^{\mathrm{c}}A_{\mathrm{s}}\right\}/\left(r_{\mathrm{a}}^{\mathrm{a}}+r_{\mathrm{a}}^{\mathrm{c}}\right)}{\Delta+\gamma\left\{1+r_{\mathrm{s}}^{\mathrm{c}}/\left(r_{\mathrm{a}}^{\mathrm{a}}+r_{\mathrm{a}}^{\mathrm{c}}\right)\right\}} \tag{10}$$

where the available energy input above the soil surface is defined as:

$$A_{\mathrm{s}}=R_{\mathrm{ns}}-G \tag{11}$$

R_{ns} can be calculated by using the Beer's law relationship:

$$R_{\mathrm{ns}}=R_{\mathrm{n}}\exp\left(-K_{\mathrm{A}}\mathrm{LAI}\right) \tag{12}$$

The two coefficients C_s and C_c are obtained as follows:

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences
Open Access
Discussions

$$C_{\mathrm{s}} = \left\{1 + R_{\mathrm{s}} R_{\mathrm{a}} / R_{\mathrm{c}} \left(R_{\mathrm{s}} + R_{\mathrm{a}}\right)\right\}^{-1} \tag{13}$$

$$C_{\mathrm{c}} = \left\{1 + R_{\mathrm{c}} R_{\mathrm{a}} / R_{\mathrm{s}} \left(R_{\mathrm{c}} + R_{\mathrm{a}}\right)\right\}^{-1} \tag{14}$$

where R_c, R_a, and R_s are given as:

$$R_{\mathrm{a}} = \left(\Delta + \gamma\right) r_{\mathrm{a}}^{\mathrm{a}} \tag{15}$$

$$R_{\mathrm{s}} = \left(\Delta + \gamma\right) r_{\mathrm{a}}^{\mathrm{s}} + \gamma r_{\mathrm{s}}^{\mathrm{s}} \tag{16}$$

$$R_{\mathrm{c}} = \left(\Delta + \gamma\right) r_{\mathrm{a}}^{\mathrm{c}} + \gamma r_{\mathrm{s}}^{\mathrm{c}} \tag{17}$$

Soil surface resistance is expressed as:

$$r_{\mathrm{s}}^{\mathrm{s}} = \exp(b_1 - b_2 \frac{\theta}{\theta_{\mathrm{s}}}) \tag{18}$$

In this study, we consider the reciprocal of bulk stomatal resistance, known as canopy conductance. The calculation of g_s^c is the same as in the PM model. The two aerodynamic resistances (r_a^a and r_a^s) and the boundary layer resistance (rac) are modeled following the approach proposed by Shuttleworth and Gurney (1990).

2.3.3. Priestley–Taylor and Flint-Childs (PT-FC) model

The Priestley–Taylor (Priestley and Taylor, 1972) model was introduced to estimate evaporation from an extensive wet surface under conditions of minimum advection (Stannard, 1993; Sumner and Jacobs, 2005). It is expressed as:

$$\lambda ET = \alpha_{PT} \frac{\Delta}{\Delta + \gamma}\left(R_n - G\right) \tag{19}$$

where α_{PT} is a unitless coefficient. The Priestley–Taylor model was modified by Flint and Childs (1991) to scale the Priestley–Taylor potential ET to actual ET for nonpotential conditions (hereafter the PT-FC model):

$$\lambda ET = \alpha \frac{\Delta}{\Delta + \gamma}\left(R_n - G\right) \tag{20}$$

where α is as a function of the environmental variables, which could be related to any process that limits ET (e.g., soil hydraulic resistance, aerodynamic resistance, stomatal resistance); however, only soil moisture status was considered to simplify ET estimation in the PT-FC model (Flint and Childs, 1991). In this model, α is defined as:

$$\alpha = \beta_1\left[1 - \exp\left(-\beta_2 \Theta\right)\right] \tag{21}$$

where Θ is calculated as

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} \tag{22}$$

2.3.4. Advection-aridity (AA) model

The AA model was first proposed by Brutsaert and Stricker (1979) and further improved by Parlange and Katul (1992). The model relies on the feedback between actual (*λET)* and potential *ET*, which assumes that actual potential *ET* should converge to wet surface *ET* at wet surface conditions. Its general form is:

$$\lambda ET = \left(2\alpha_{PT} - 1\right)\frac{\Delta}{\Delta + \gamma}\left(R_n - G\right) - \frac{\gamma}{\Delta + \gamma}\frac{\rho\left(q^* - q\right)}{r_a} \tag{23}$$

where α_{PT} is the Priestley–Taylor coefficient, usually taken as 1.26 (Priestley and Taylor, 1972); and r_a is similar to that used for the Penman-Monteith model (Brutsaert and Stricker, 1979; Brutsaert, 2005; Ershadi

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences
Open Access
Discussions

et al., 2014). This model is based mainly on meteorological variables and does not require any information related to soil moisture, canopy resistance or other measures of aridity (Ershadi et al., 2014). In this study, similar to the PT-FC model, we modified α_{PT} to α, which is calculated using the same equation as in the PT-FC model. The detailed list of symbols and physical characteristics in ET models are stated in Appendix A.

2.4 BME Estimation

2.4.1 Thermodynamic Integration Estimator

Estimating the BME using power posterior estimators such as thermodynamic integration (TI) (Lartillot and Philippe, 2006) depends mainly on the likelihood $p(\mathbf{D}|\boldsymbol{\theta},M)$ calculation. The main idea of power posterior sampling is to define a path that links the prior to the unnormalized posterior. Thus, using an unnormalized power posterior density

$$q_\beta(\boldsymbol{\theta}) = p(\mathbf{D}|\boldsymbol{\theta},M)^\beta p(\boldsymbol{\theta}|M) \tag{24}$$

the power coefficient $\beta \in [0,1]$ is a scalar parameter for discretizing a continuous and differentiable path linking two unnormalized power posterior densities. The unnormalized power posterior density $q_\beta(\boldsymbol{\theta})$ in Equation (24) uses the normalizing constant Z_β to yield the normalized power posterior density:

$$p_\beta(\boldsymbol{\theta}) = \frac{q_\beta(\boldsymbol{\theta})}{Z_\beta} \tag{25}$$

such that

$$Z_\beta = \int q_\beta(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{26}$$

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

The above integral takes a simplified form by the potential:

$$U(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial \beta} \tag{27}$$

thus, the integration can be directly estimated by following:

$$p(\mathbf{D}|M) = \frac{Z_1}{Z_0} = \exp\left\{\int_0^1 E_\theta\left[\ln p(\mathbf{D}|\boldsymbol{\theta}, M)\right] d\beta\right\} \tag{28}$$

2.4.2 Power posterior sampling

The Metropolis acceptance ratio is $\alpha_k = \min\left(1, [\alpha_{k,power-postrior}\alpha_{k,prior}]\right)$ with the power posterior ratio given by $\alpha_{k,power-posterior} = \left(\alpha_{k,posterior}\right)^{\beta_k}$. The prior probability ratio $\alpha_{k,prior} = \Pr(\boldsymbol{\theta}_{k,new} \mid M) / \Pr(\boldsymbol{\theta}_{k,old} \mid M)$ is the ratio of the probability of the newly proposed sample $\boldsymbol{\theta}_{k,new}$ and the probability of the previously accepted sample $\boldsymbol{\theta}_{old}$. The posterior probability ratio $\alpha_{k,posterior} = L(\mathbf{D} \mid \boldsymbol{\theta}_{k,new}, M) / L(\mathbf{D} \mid \boldsymbol{\theta}_{k,old}, M)$ is the likelihood ratio of samples $\boldsymbol{\theta}_{k,new}$ and $\boldsymbol{\theta}_{k,old}$, and β_k is the power posterior coefficient. Thus, to use the DREAM to sample any power posterior distributions (Bayesian inference and the DREAM algorithm please see the Appendix B), the regular Metropolis acceptance ratio $\alpha = min\left(1, [\alpha_{posterior}\alpha_{prior}]\right)$ is changed to $\alpha_k = \min\left(1, [\alpha_{k,power-postrior}\alpha_{k,prior}]\right)$ in DREAM.

Since there has been no theoretical method so far for selecting β values (Liu et al., 2016), we determined these values using an empirical but straightforward method. Following Xie et al. (2011), a schedule of the power posterior coefficients β_k is generated by

$$\beta_k = (k / K)^{1/\alpha} \tag{29}$$

for k =0, 1, 2…, K. Using $\alpha = 0.3$ and $K = 20$ is a reasonable initial choice.

2.4.3 Estimation of BME using TI

After obtaining the power posterior samples, the corresponding likelihoods are used to estimate the BME. This step simply requires the log-likelihoods $\log p(\mathbf{D}|\boldsymbol{\theta}_{k,i}, M)$ to be inserted into the following equation:

$$\hat{r}_{TI} \equiv p(\mathbf{D} \mid M) = \exp\left(\int_0^1 y_\beta d\beta\right) = \exp\left(\sum_{k=0}^{K} r_{TI,k}\right) \tag{30}$$

such that

$$r_{TI,k} = (\beta_k - \beta_{k-1})\left[\frac{y_k - y_{k-1}}{2}\right] \tag{31}$$

and

$$y_k = E_\beta[\log p(\mathbf{D} \mid \boldsymbol{\theta}_k, M)] = \frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{D}|\boldsymbol{\theta}_{k,i}, M) \tag{32}$$

Each panel in this one-dimensional integral is given by $r_{TI,k}$ for the case of the trapezoidal rule (Eq. 31), and the summation of these panels gives the natural logarithm of BME.

2.5 Traditional statistical evaluations

The traditional statistics for evaluating model performance include correlation-based measures of $R^2$and slope, relative error measures of IA and EF, and absolute error measures of RMSE and mean bias error (MBE) (Poblete-Echeverria and Ortega-Farias, 2009). Their definitions are as follows:

$$\mathrm{IA} = 1 - \frac{\sum_{t=1}^{n}[O(t) - M(t)]^2}{\sum_{t=1}^{n}\left[\left|O(t) - \overline{O(t)}\right| + \left|O(t) - \overline{M(t)}\right|\right]^2} \tag{33}$$

$$\mathrm{EF}=1-\frac{\sum_{t=1}^{n}\left[O(t)-M(t)\right]^2}{\sum_{t=1}^{n}\left[O(t)-\overline{O(t)}\right]^2} \tag{34}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left[O(t)-M(t)\right]^2} \tag{35}$$

$$\mathrm{MBE}=\frac{1}{n}\sum_{t=1}^{n}\left[O(t)-M(t)\right] \tag{36}$$

where $O(t)$ is the observations and $\overline{O(t)}$ is the mean observation at time *t*, $M(t)$ is the modeled value estimated by the posterior median parameter values, and *n* is the total number of the observed values.

3. Results

3.1 Parameter estimation

There were five parameters g_{max}, D_{50}, Q_{50}, K_Q and Q_A in the PM model, and two additional parameters, b_1 and b_2, in the SW model. The PT-FC and AA models include two parameters, denoted as β_1 and β_2 (Table 1). The prior probability density of each parameter was specified as a uniform distribution with the ranges listed in Table 1. A total of 50,000 realizaitions were generated with DREAM and the calibration period data were from DOY 154 to DOY 202. In the calculations, the chain number, *N*, was equal to the parameter number, i.e., $N = 5$, 7, 2 and 2 for the PM, SW, PT-FC and AA models, respectively. For each model, the first 10,000 samples were discarded as burn-in data, and the remaining 40,000 samples were used to set up posterior density functions for each chain.

To understand the efficiency and convergence of DREAM for the ET models, Figure 1 shows the trace plots of the G-R statistic for each of the different parameters in the PM and SW models with the different color (PT-FC and AA models not shown). The algorithm requires about 8,000 generations to make the G-R statistic smaller than 1.2 for the both models. Obviously, the complete mixing of the different chains and convergence of DREAM were attained after about 620 and 450 generations for PM and SW models, respectively. The acceptance rates for the PM and SW models were about 15.3% and 18.9%, respectively.

Histograms and cumulative distribution functions (CDFs) of the DREAM-derived marginal distributions of the parameters are presented in Figure 2 and summarized in Table 2 by Maximum Likelihood Estimates (MLEs), posterior medians and 95% probability intervals. The uppercase in Figures 2A-2E, 2I-2O, 2F-2G, and 2H and 2P showed histograms, and the corresponding lowercase of 2a-2e, 2i-2o, 2f-2g, and 2h and 2p show CDFs, for the PM, SW, PT-FC and AA models, respectively.

Parameter g_{max} (Fig. 2A) in the PM model, parameters g_{max}, K_A, b_1, b_2 (Fig. 2I, 2M, 2N, 2O) in the SW model, and parameter β_1 (Fig. 2F) were well constrained and occupied a relatively small range. These parameter sample displayed a unimodal distribution, and became approximately symmetric. Parameters Q_{50}, D_{50}, K_Q and Q_A (Fig. 2B-2E) in the PM model and parameters D_{50}, K_Q in the SW model (Fig. 2K-2L) exhibited relatively large uncertainty reductions. However, the histograms obviously deviated from normality and tended to concentrate in the lower bounds. When the upper limits of these parameters were decreased, similar histograms were reached (not shown) and still did not show statistically meaningful distributions. In contrast, Q_{50} was not only poorly constrained (Fig. 2J) but was also the edge-hitting parameter in the SW model. In addition, the corresponding distributions of the same parameter in different models were slightly different; for example, the mean of g_{max} in the PM model (0.04 mm s^{-1}) was less than that in the SW model (0.01 mm s^{-1}) (Fig. 2A and 2I; Table 2), except that D_{50} in the PM and SW models and β_2 in the PT-FC and AA models exhibited similar regions (Fig. 2C and 2K, 2G and 2P; Table 2).

3.2 Performance of the models

The performances of the four evaporation models were evaluated during the whole season in 2014. The calibrated parameters of the four models were used and individual ET models were run to estimate the half-hourly λET values. Statistical results for the performance of the models were summarized in tables as the regression line slope, R^2, RMSE, MBE, IA, and EF as shown in Table 3. The regressions between measured and modeled λET values and MBE are shown in Figures 3 and 4, respectively.

In general, the four models produced slightly better fits to the measured λET for all the seasons with R^2 larger than 0.75 (Fig. 3). However, obvious discrepancies among models were detected by comparing measured and modeled λET. According to the regression line slope and MBE, the PM model overestimated ET by 1% with a MBE of -9.52 W m^{-2}, and the SW overestimates ET by 5% with a relatively higher MBE of -19.07 W m^{-2} compared to the PM model. The PT-FC and AA models tended to underestimate λET by 9% and 8% with an MBE of 25.42 and 23.29 W m^{-2}, respectively. From a comparison between the slope and MBE, the PM model performance was higher than the SW, PT-FC and AA models, with a slope almost equal to 1 and with relatively lower MBE. The SW model was ranked second, while the AA model was comparable to the PT-FC, but slightly higher, and was ranked third. However, if R^2, RMSE, IA, and EF were used to evaluate the model performances, the SW model had the best overall performance with R^2=0.83, RMSE= 76.34 W m^{-2}, IA = 0.95 and EF = 0.79. The second-best model was the PM model with R^2 = 0.76, RMSE = 85.38 W m^{-2}, IA = 0.93 and EF = 0.74.The PT-FC performance was ranked third with R^2 = 0.75, RMSE = 94.39 W m^{-2}, IA = 0.92 and EF = 0.68, while the

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

AA model ranked fourth with $R^2 = 0.75$, RMSE = 95.09 W m^{-2}, IA = 0.92 and EF = 0.67. Based on the analysis of these traditional statistical criteria, the performances of the PT-FC and AA models yielded similar results. The observed and modeled λET for the four ET models were tightly grouped along the regression lines (Figure 3), and the PT-FC and AA models had similar modeled ET values with a similar degree of point scattering along the regression lines (Figure 3c-3d).

Figure 4 shows that large seasonal variations were exist in MBE for the four ET models. From the variations of the MBE, the estimated λET values for all models were generally lower than the measured values before the early jointing stage of maize growth (DOY 154-177, left dashed line) and after the late maturity stage (DOY 256-265, right dash line) with the corresponding LAI < 2.5 m^2 m^{-2}. More positive MBE values for the PT-FC and AA models after the late maturity stage indicated their underestimated performances; however, these estimations appeared even more consistent with a symmetrical scattering of points along the 0-0 line (Figure 4c, 4d) during DOY 177-256 with LAI > 2.5 m^2 m^{-2}.

3.3 Comparison of the models using BME

Since there was no theoretical method so far for selecting β values, we determined these values using empirical but straightforward methods. For any different power coefficient of $\beta \in [0,1]$, a sample was drawn from the distribution p_β (Eq. 25) through running DREAM. Figure 5 showed the evolution of *ln* $p(D|\theta, M)$ for the four models as a function of β for a dataset covering the entire period. The potential values of the PM model increased from -6533.02 (the logarithm of the prior likelihood) to -6290.71, and the potential values increased from -6544.49 to -6016.17 for the SW model. In addition, the potential values increased from -6708.02 to -6361.76 for the PT-FC model and from -7732.98 to -7033.32 for the AA model. Table 3 showed that the estimated BME is -6300.5 natural log units (nits) for the PM model,

-6025.1 nits for the SW model, -6366.8 nits for the PT-FC model, and -7042.8 nits for the AA model. The BME for the SW model was substantially larger than that for the other three models, and the BME for the AA model was the smallest. Although the parameters of the PM model were less than for the SW model, the potential evolution of the SW model was substantially different to that of the PM model. In summary, the PT-FC and AA models, consisting of the same number of parameters, had similar potential patterns of evolution with the coefficient β_k. Although adding more β_k values may improve the BME estimation, it was not undertaken because of the computational cost. For each β_k value, 150,000 DREAM simulations were large enough to ensure convergence.

4. Discussion

4.1 Parameter uncertainty analysis

With regard to the efficiency of the DREAM algorithm, the acceptance rates of the PM (15.33%) and SW (18.94%) models were much higher than some MCMC algorithms which used in the previous studies, like 0.019% in the population Monte Carlo sampling algorithm (Sadegh et al., 2014). This was a large improvement in search efficiency, which in large part resulted from its ability to sample groups of variable in turn. Furthermore, this method ran multiple chains in parallel and adaptively updated the scale and orientation of the proposal distribution (Vrugt et al., 2008). Therefore, the DREAM scheme substantially improved not only the convergence, but also its sampling efficiency for ET models.

The results showed that the DREAM algorithm successfully reduced the assumed prior uncertainties from the large number of parameters in the four models. The well-constrained parameters were those that had significant contribution. For example, the ecophysiological parameter g_{max}, in both the PM and SW models, the maximum stomatal conductance of leaves, and the soil surface resistance parameters b_1 and b_2 in the SW models, all had large influences on the evaluated ET. Thus, their effects were relatively independent compared to the other meteorological parameters in the models. The posterior mean value of g_{max} (0.04 m s^{-1}) in the PM model from our study was close to that (0.05 m s^{-1}) reported in northwestern China (Li et al., 2013; Zhu et al., 2014), but g_{max} (0.01 m s^{-1}) in the SW model was less than the reported value. The estimated posterior means for b_1 and b_2 were different (b_1 = 9.3, b_2 = 6.2) from those for maize suggested by Zhu et al. (2014) using the same equation of soil surface resistance (r_s^s). Though Zhu et al. (2014) concluded that the responses of g_s^c to VPD and LAI calculated using the modified Leuning model were close to those using Javis model (Jarvis, 1976), Li et al. (2015) showed that the performance of PM model was different using the two canopy resistance formula. Therefore, the different results of parameters b_1 and b_2 between our study and the previous study by Zhu et al. (2014) were mainly due to the usage ofdifferent canopy conductance models.

For edge-hitting parameters, their uncertianties may be the outcome of model biases or EC-measured ET data, or the characteristic time scale of parameters govern processes that was not exactly on the order of half-hours (Braswell et al., 2005). For example, Q_{50} and D_{50} govern changes in visible radiation flux and humidity deficit at which stomatal conductance at its half maximum value, which may change over a shorter or longer time scale rather than half-hours. K_Q was another parameter that cannot be well constrained, and this may be resulted from either the estimated ET was insensitive to these parameters, or there were correlations between the parameters. We expected a complementary correlative relationship between the visible radiation flux and extinction coefficient for shortwave radiation, which indicated that the information in EC-measured ET data was insufficient to separate these parameters, and therefore the

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

parameters cannot be constrained separately.

The sensitive parameters (such as g_{max}, b_1 and b_2) were just corresponding to the well-constrained parameters. Therefore, the major parameters in PM and SW models were well optimized, except that several parameters (Q_{50} and K_Q) appeared to be not well constrained. In addition, the posterior parameter bounds exhibited a larger reduction using the DREAM algorithm compared with other studies using the Metropolis–Hasting algorithm (Zhu et al., 2014). This further demonstrated that DREAM can efficiently handle problems involving high-dimensionality, multimodality, nonlinearity, and local optima.

In general, parameters related to soil surface resistance in the SW model were well evaluated, while parameters related to canopy surface resistance in PM and SW models were poorly estimated. Therefore, using a reliable canopy surface resistance equation in the ET model was crucial for improving its performance. In addition, in our study, the traditional approach was used to quantify the uncertainty which assumed that the uncertainty mainly came from the parameter uncertainty. However, this method did not explicitly consider errors in the input data and model structural inadequacies. This was unrealistic for real applications, and it was desirable to develop a more reliable inference method to treat all sources of uncertainty separately and appropriately (Vrugt et al., 2008). Moreover, simultaneous direct measurement by micro-lysimeter of sap flow and daily soil evaporation will further help to constrain the model parameters.

4.2 Evaluation and selection of the models

In this study, the traditional statistical measures and BME were chosen to evaluate and compare the performance of four ET models. From the respective composition of these measures, the statistical measures can be divided into residual-based measures (such as regression slope and MBE) and squared-residual-based measures (such as R^2, RMSE, IA, and EF). Table 3 shows the values evaluated by BME method, residual-based and squared-residual-based measures. By comparison, the estimates obtained within the same measure (residual-based or squared-residual-based) were congruent. For example, slope and MBE have similar results in the residual-based measures. However, the results from different kind of measures were incongruent; for example, PM model outperformed SW model according to the residual-based measures, but PM model was worse than SW model based on the squared-residual-based measures. The comparative analysis showed a consistency between BME and the squared-residual-based statistics, whereas residual-based criteria were obvious disagreement with the BME measures. It revealed that the more complex SW model was the best model based on the BME and squared-residual-based statistics. The rank order of overall performance of the models from best to worst was: SW, PM, PT-FC, and AA model.

Previous studies had shown that BME evaluated by the TI provided estimates similar to the true values and selected the true model if the true model was included within the candidate models (Marshall et al., 2005; Lartillot and Philippe, 2006). Meanwhile, some argured that Bayesian analysis would choose the simplest model (Jefferys and Berger, 1992; Xie et al., 2011) because of the best trade-off between good fit with data and model complexity (Schöniger et al., 2014). In this case, the most complex SW model had the highest BME and was chosen as the best-behaved model. This likely resulted from the fact that the complex SW model was indeed the most reliable model among the alternative ET models. SW model was a two-layer model, and estimated soil evaporation and plant transpiration separately, but PM model was a single-layer model while the plant transpiration and soil evaporation cannot be separated (Monteith, 1965).

The PT-FC model was a simplified model of PM, and it only required meteorological and radiation information (Priestley and Taylor, 1972), whereas AA model only relied on the feedback between actual ET and potential ET (Brutsaert and Stricker, 1979). Based on these physical mechanisms and processes for these ET models, the rank order of the models was reasonable.

The estimates showed that the maximum values of R^2, IA and EF, and the minimum value of RMSE, all selected the most complex SW model as the best performing model. The results indicated that the SW model was the best performing model evaluated by squared-residual-based measures, which resulted from the ability of the model to fit the measured data, irrespective of model complexity. It was interesting to note that both the squared-residual-based measures and the BME consistently yielded the same rank order. Although the squared-residual-based measures seemed to identify a reasonable rank order, this had often not been the case, since the simple traditional statistical measures were known to usually provide a biased view of the efficacy of a model (Kessler and Neas, 1994; Legates and McCabe, 1999). In addition, sensitivity to outliers was associated with these measures and leads to relatively high values due to the squaring of the residual terms (Willmott, 1981). Furthermore, these traditional statistical measures ignored the priors, without penalizing model complexity, which was in fact used in a Bayesian analysis. The dimensionality (model's parameter space) not only affected model evaluation by BME (Schöniger et al., 2014) but it may also affect the evaluation using traditional statistical measures. Here, two-dimensional models of PT-FC and AA provided identical estimates of R^2 and IA. This was most likely because both the PT-FC and AA models had the same dimensions and a similar model structure, whereas BME estimates remain well-behaved for the two ET models. Marshall et al. (2005) argued that EF would provide an incorrect conclusion, and Samani et al. (2018) suggested that RMSE also selected the complex model as the best performing model. Thus, we deduced that SRB measures are also problematic. As for slope and MBE, these residual-based measures were obvious disagreement with the BME measure. Part of the lower values of slope and MBE may be counter balanced by the higher values of slope and MBE, thus these criterias provided an erroneous and unreliable model evaluation. Therefore, the squared-residual-based and residual-based measures were not certain to provide reasonable results in terms of model ranking.

4.3 Analysis of model-data mismatch

Conceptual and structural inadequacies of the hydrological model and measurement errors of the model input (forcing) and output (calibration) data introduced errors in the estimated parameters and model simulations (Laloy, 2014). Hydrological systems were indeed heavily input driven and errors in forcing data can dramatically impair the quality of calibration results and model output (Bardossy and Das, 2008; Giudice, 2015). Measurement errors were raised for a variety of reasons, including unreasonable gap-filling in rainy days; dew and fog; inadequate areal coverage of point-scale soil water measurement; mechanical limitations of the EC system; and inaccurate measurements of wind-speed, soil water, radiation and vapor pressure deficit. ET processe was described using equations that can only capture parts of the complex natural processes and the model structures were an inherent simplification of the real system. These inadequacies can thus lead to biased parameters and implausible predictions.

In our study, the results indicated that the PM and SW models overestimated the half-hourly ET compared to the measured ET. Several studies also indicated that the ET values were overestimated by the PM model (Fisher et al., 2005; Ortega-Farias et al., 2006; Li et al., 2015) and the SW model (Li et al., 2013; Li et al., 2015; Zhang et al., 2008). Possible reasons for the inaccurate estimates included the following: (1) Anisotropic turbulence with weak vertical and strong horizontal fluctuation leads to energy imbalance. The total turbulent heat flux was lower by ~10–30% compared to the available energy in many land surface experiments (Tsvang et al., 1991; Beyrich et al., 2002; Oncley et al., 2007; Foken et al., 2010) and influx networks (Franssen et al., 2010). Liang et al. (2017) also showed an energy imbalance result in the semiarid area in China, and indicated that the energy balance closure ratio ranged from 0.52 to 0.90 during the daytime, whereas it was about 0.25 during night-time. However, the measured ET only included vertical flux and not horizontal flux, leading to the measured ET being lower than that of modeled ET by the PM and SW models using the available energy. (2) The absence of a mechanistic representation of the physiological response to plant hydrodynamics cause it difficult for the available ET models to resolve the dynamics of intradaily hysteresis, producing patterns of diurnal error, while the imbalance or lack of between-leaf water demand and soil water supply imposes hydrodynamic limitations on stomatal conductance (Thomsen et al., 2013; Zhang et al., 2014; Matheny et al., 2014). Li et al. (2015) also concluded that neglecting the restrictive effect of the soil on water transport in empirical canopy resistance equations can result in large errors in the partial canopy stage. However, these equations can simulate ET accurately under the full canopy stage (Alves and Pereira, 2000; Katerji and Rana, 2006; Katerji et al., 2011; Rana et al., 2011). Li et al. (2015) showed the PM model combined with the canopy resistance overestimated maize ET during the partial and dense canopy stages by 16% and 13%, respectively (Leuning, 2008). Moreover, the PM model coupled with the canopy resistance overestimated vineyard ET during the entire growth stage by 29% (Leuning, 2008).

The estimated ET for the PT-FC and AA models was generally lower than the measured values during the entire season. In addition, the four models also underestimated the ET during periods of partial cover (LAI < 2.5 m^2 m^{-2}). Especially during the late maturity stage, the PT-FC and AA models consistently underestimated ET and provided the worst simulated ET. The underestimation probably resulted from the following: (1) Non classical situations, such as the oasis effect, may occur in the study area. Strong evaporation from the moist ground and plants results in latent heat cooling. However, this upward latent heat flux was opposed by a downward sensible heat flux from the warm air to the cool ground, and thus the latent heat flux was positive while the sensible heat flux is negative. Therefore, the latent heat flux can be greater in magnitude than the solar heating, because of the additional energy extracted from the warm air by evaporation (Stull, 1988). (2) Lack of mechanistic representation of rainfall interception in ET models probably also led to inaccurate simulation on shortly after rainy days. Bohn and Vivoni (2016) found that evaporation of canopy interception accounted for 8% of the annual ET across the North American monsoon region.

Comparing the AA and PT-FC models, the former included forcing data of available radiation, soil water content and relative humidity, but the PT-FC model only requires available radiation and soil water content and was independent of relative humidity. However, the similar statistical results and similar degrees of MBE scatter indicated that relative humidity has little influence on the AA model simulation. The consistent and consecutive underestimation of ET by the PT-FC and AA models during the late maturity stage showed that the model-data disagreement is caused mainly by regional advection and rainfall interception, because atmospheric processes and thermally-induced circulation can only occur at certain times and during certain days. Therefore, we suggested that the consistent underestimation of ET by the PT-FC and AA models primarily results from conceptual and structural inadequacies, energy imbalance, and soil water stress. Although the PM and SW models shared a common theoretical basis and the PT-FC model was the simplification of the PM model, these models performed significantly differently.

Part of the overestimation of ET by the PM and SW models, caused by coupling with the canopy resistance, may be offset by underestimation caused by energy imbalance and soil water stress. However, underestimation of ET by the PT-FC and AA models cannot be counterbalanced by overestimation during the later maturity stage because the PT-FC and AA models are independent of the canopy resistance. Consequently, the half-hourly patterns of errors in the prediction of ET by the PM and SW models were characterized by symmetry and a low degree of scatter, but the PT-FC and AA models exhibited consistent and asymmetrical error patterns.

By contrast, other studies showed that the PM model (Kato et al., 2004) and the SW model (Chen et al., 2015) underestimated half-hourly ET. As for the PT-FC and AA models, while some studies reported that the PT-JPL (Zhang et al., 2017) and AA model showed an overall poor performance, however, other studies have indicated that the AA method performed well for both maize and canola crops (Liu et al., 2012). Therefore, the performance of the four ET models appears to vary not only for different crops and locations (Zhu et al., 2014) but also for different meteorological, physiological and soil conditions. Moreover, the performance was also related to the stage of crop growth.

5. Conclusions

This study illustrated the use of the Bayesian approach for the statistical analysis and model selection of four widely used ET models. BME can be used to rank the alternative models in our study, although numerical evaluation of BME is computationally expensive particularly for high-dimensional models. Bayesian model comparison identified the SW model as the best ET model. Although the squared-residual-based measures, including R^2, IA, RMSE, and EF, provide a congruent model ranking with that of BME, it must be noted that these squared-residual-based measures do not allow using prior information for comparing the models. We advocated that caution is needed when using these statistical methods, and that BME should be used instead. In contrast, residual-based measures disagree with the BME measure, and thus these measures can not be used for evaluating model performance.

The model–data mismatches were analyzed to facilitate model improvement after using Bayesian model calibration and comparison. The results indicated that model–data mismatches are mainly resulted from energy imbalance caused by anisotropic turbulence, the absence of a mechanistic representation of the physiological response to plant hydrodynamics, and additional energy induced by advection processes. Among them, energy imbalances and additional energy were related to forcing data error rather than to an unreasonable model structure. Thus, understanding the process of the physiological response to plant hydrodynamics, such as developing or selecting more reasonable and process-based canopy resistance models, was essential for improving the performance of evapotranspiration models. Overall, in our study, the applications of Bayesian calibration, Bayesian model evaluation and analysis of model–data mismatches, provided a promising framework for reducing uncertainty and improving the performance of ET models.

Author contribution

Guoxiao Wei and Xiaoying Zhang designed the experiments. Ning Yue and Fei Kan carried them out. Ming Ye developed the model selection scheme. Guoxiao Wei performed the simulations. Guoxiao Wei and Xiaoying Zhang prepared the manuscript with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

We thank Ying Guo, Huihui Dang, Jun Dong for the data collection and analysis. This work was funded by the National Natural Science Foundation of China (Nos. 41471023). The third author was supported in part by DOE Early Career Award DE-SC0008272 and National Science Foundation-Division of Earth Science Grant 1552329. All observed data used in this study are from Heihe Watershed Allied Telemetry Experimental Research (HiWATER). We thank all the staff who participated in HiWATER field campaigns. Considerate and helpful comments by anonymous reviewers have considerably improved the manuscript.

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle, in: Breakthroughs in Statistics, vol. 1, Foundations and Basic Theory, edited by: Kotz, S. and John-15 son, N. L., Springer-Verlag, New York, USA, 610–624, 1973.

Allen, R. G., Perista, L. S., Raes, D., and Smith, M.: Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements; FAO Irrigation and Drainage apers-56, FAO-Food and Agriculture Organization of the United Nations, Rome, 1998.

Alves, I. and Pereira, L. S.: Modeling surface resistance from climatic variables? Agric. Water Manage., 42, 371–385, 2000.

Aubinet, M., Grelle, A., Ibrom, A., Rannik, Ü., Moncrieff, J., and Foken, T.: Estimates of the annual net carbon and water exchange of forests: the euroflux methodology, Adv. Ecol. Res., 30(1), 113-175, 1999.

Baldocchi, D. D.: Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future, Glob. Change. Biol., 9, 479–492, 2003.

Bardossy, A., and Das, T.: Influence of rainfall observation network on model calibration and application, Hydrol. Earth Syst. Sci., 12(1), 77-89, https://doi:10.5194/hess-12-77-2008, 2008.

Barton. I. J.: A Parameterization of the Evaporation from Nonsaturated Surfaces, J. Appl Meteorol., 18(1), 43-47, 1979.

Beyrich, F., Richter, S. H., Weisensee, U., Kohsiek, W., Lohse, H., de Bruin, H. A. R., Foken, T., Göckede, M., Berger, F., Vogt, R., and Batchvarova, E.: Experimental determination of turbulent fluxes over the heterogeneous litfass area: selected results from the litfass-98 experiment, Theor. Appl. Climatol., 73(1–2), 19–34, https://doi:10.1007/s00704-002-0691-7, 2002.

Bohn, T. J., and Vivoni, E. R.: Process-based characterization of evapotranspiration sources over the North American monsoon region, Water Resour. Res., 52, 358–384, https://doi:10.1002/2015WR017934, 2016.

Bonan, G.: Ecological climatology: concepts and applications, Cambridge University Press, 2008.

Braswell, B. H., Sacks, W. J., Linder, E., and Schimel, D. S.: Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations, Global. Change. Biol., 11, 335–355, 2005,

Brutsaert, W., and Han, S.: An advection-aridity approach to estimate actual regional evapotranspiration, Water Resour. Res., 15(2), 443-450, 1979.

Brutsaert, W.: Hydrology: An Introduction. Cambridge University Press, Cambridge, 2005.

Chen, D. Y., Wang, X., Liu, S. Y., Wang, Y. K., Gao, Z. Y., Zhang, L .L., Wei, X. G., and Wei, X. D.: Using

Bayesian analysis to compare the performance of three evapotranspiration models for rainfed jujube (Ziziphus jujuba Mill.) plantations in the Loess Plateau, Agr. Water. Manage., 159, 341–357, 2015.

Elshall, A. S., Ye, M., Pei, Y., Zhang, F., Niu, G. Y., and Barron-Gafford, G. A.: Relative model score: A scoring rule for evaluating ensemble simulations with application to microbial soil respiration modeling, Stoch. Env. Res. A., 1-11, DOI: 10.1007/s00477-018-1592-3, 2018.

Ershadi, A., Mccabe, M. F., Evans, J. P., Chaney, N. W., and Wood, E. F.: Multi-site evaluation of terrestrial evaporation models using fluxnet data, Agric. For. Meteorol., 187(8), 46-61, 2014.

Ershadi, A., McCabe, M .F., Evans, J. P., and Wood, E. F.: Impact of model structure and parameterization on Penman–Monteith type evaporation models, J. Hydrol., 525, 521–535, 2015.

Fisher, J. B., DeBiase, T. A., Qi, Y., Xu, M., and Goldstein, A. H.: Evapotranspiration models compared on a Sierra Nevada forest ecosystem, Environ. Model. Softw., 20 (6), 783–796, 2005.

Flint A. L., Childs, S. W.: Use of the Priestley–Taylor evaporation equation for soil water limited conditions in a small forest clearcut, Agric. For. Meteorol., 56(3–4), 247–260, 1991.

Foken, T., Mauder, M., Liebethal, C., Wimmer, F., Beyrich, F., Leps, J. P., Raasch, S., DeBruin, H. A. R., Meijninger, W. M. L., and Bange, J.: Energy balance closure for the LITFASS-2003 experiment, Theor. Appl. Climatol., 101(1–2), 149–160, https://doi:10.1007/s00704-009-0216-8, 2010.

Franssen, H. J. H., Stöckli, R., Lehner, I., Rotenberg, E., and Seneviratne S. I.: Energy balance closure of eddy-covariance data: A multisite analysis for European FLUXNET stations, Agric. For. Meteorol., 150(12), 1553–1567, https://doi:10.1016/j.agrformet.2010.08.005, 2010.

Gelman, A., and Rubin, D. B.: Inference from iterative simulation using multiple sequences, Stat. Sci., 7, 457– 472, 1992.

Gelman, A.: Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, Stat. Sci. 13, 163–185, 1998.

Giudice, D., Albert, C., Rieckermann, J., and Reichert, P.: Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation, Water Resour. Res., 52, 3162–3186, https://doi:10.1002/2015WR017871, 2016.

Jarvis, P. G.: The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. Philos, T. Roy. Soc. B., 273, 563–610, 1976.

Jefferys, W. H., and Berger, J. O.: Sharpening Ockham's razor on a Bayesian strop, Am. Sci., 89, 64-72, 1992.

Kashyap, R. L.: Optimal choice of AR and MA parts in autoregressive moving average models, IEEE Trans. Pattern Anal. Mach. Intell., 4(2), 99–104, 1982.

Katerji, N., and Rana, G.: Modelling evapotranspiration of six irrigated crops under Mediterranean climate conditions, Agric. For. Meteorol., 138, 142–155, 2006.

Katerji, N., Rana, G., Fahed, S.: Parameterizing canopy resistance using mechanistic and semi-empirical estimates of hourly evapotranspiration: critical evaluation for irrigated crops in the Mediterranean, Hydrol. Process., 25, 117–129, 2011.

Kato, T., Kimura, R., and Kamichika, M.: Estimation of evapotranspiration, transpiration ratio and water-use efficiency from a sparse canopy using a compartment model, Agric. Water Manage., 65, 173–191, 2004.

Kessler, E., and Neas, B.: On correlation, with applications to the radar and raingage measurement of rainfall, Atmos. Res., 34, 217-229, 1994.

Laloy, E., Linde, N., Jacques, D., and Vrugt, J. A.: Probabilistic inference of multi-Gaussian fields from indirect hydrological data using circulant embedding and dimensionality reduction, Water Resour. Res., 51, 4224–4243, https://doi:10.1002/2014WR016395, 2015.

Lartillot, N., and Philippe, H.: Computing Bayes factors using thermodynamic integration, Syst. Biol., 55(2), 195-207, 2006.

Legates, D. R., and Davis, R. E.: The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches, Geophys. Res. Lett., 24, 2319-2322, 1997.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences
Open Access
Discussions

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

Leuning, R., Zhang, Y. Q., Rajaud, A., Cleugh, H., and Tu, K.: A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman–Monteith equation, Water Resour. Res., 44, W10419, http://dx.doi.org/10.1029/2007WR006562, 2008.

Liang, J., Zhang, L., Cao, X., Wen, J., Wang, J., and Wang, G.: Energy balance in the semiarid area of the Loess Plateau, China, J. Geophys. Res. Atmos., 122, 2155–2168, https://doi:10.1002/2015JD024572, 2017.

Li, S., Kang, S., Zhang, L., Ortega-Farias, S., Li, F., Du, T., Tong, L., Wang, S., Ingman, M., and Guo, W.: Measuring and modeling maize evapotranspiration under plastic film-mulching condition, J. Hydrol., 503, 153–168, 2013.

Li, S., Zhang, L., Kang, S., Tong, L., Du, T., Hao, X., Zhao, P., 2015. Comparison of several surface resistance models for estimating crop evapotranspiration over the entire growing season in arid regions. Agric. For. Meteorol. 208, 1-15

Li, X., Cheng, G. D., Liu, S. M., Xiao, Q., Ma, M. G., Jin, R., Che, T., Liu, Q. H., Wang, W. Z., Qi, Y., Wen, J. G., Li, H. Y., Zhu, G. F., Guo, J. W., Ran, Y. H., Wang, S. G., Zhu, Z. L., Zhou, J., Hu, X. L., and Xu, Z. W.: Heihe Watershed Allied Telemetry Experimental Research (HiWATER): Scientific objectives and experimental design, B. Am. Meteorol. Soc., 94, 1145–1160, 2013.

Liu, S. M., Xu, Z. W.,Wang,W. Z., Jia, Z. Z., Zhu, M. J., Bai, J., and Wang, J. M.: A comparison of eddy-covariance and large aperture scintillometer measurements with respect to the energy balanceclosure problem, Hydrol. Earth Syst. Sci., 15, 1291–1306, https://doi:10.5194/hess-15-1291-2011, 2011.

Liu, G., Liu, Y., Hafeez, M., Xu, D., Vote, C.: Comparison of two methods to derive time series of actual evapotranspiration using eddy covariance measurements in the southeastern Australia, J. Hydrol., 454–455 (4), 1–6, 2012.

Liu, P., Elshall, A. S., Ye, M., Beerli, P., Zeng, X., Lu, D., and Tao, Y.: Evaluating marginal likelihood with

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences
Open Access
Discussions

thermodynamic integration method and comparison with several other numerical methods. Water Resour. Res., 52(2), 734-758, https://doi:10.1002/2014WR016718, 2016.

Marshall, L., Nott, D., and Sharma, A.: Hydrological model selection: A Bayesian alternative, Water Resour. Res., 41(10), 3092-3100, https://doi: 10.1029/2004WR003719, 2005.

Matheny, A. M., Bohrer, G., Stoy, P. C., Baker, I. T., Black, A. T., Desai, A. R., Dietze, M. C., Gough, C. M., Ivanov, V. Y., Jassal, R. S., Novick, K .A., Schäfer, K. V. R., and Verbeeck, H.: Characterizing the diurnal patterns of errors in the prediction of evapotranspiration by several land-surface models: An NACP analysis, J. Geophys. Res. Biogeosci., 119(7), 1458-1473, 2014.

Monteith, J. L.: Evaporation and environment, Symp. Soc. Exp. Biol., 19, 205–234, 1965.

Morison, J. I. L., Baker, N. R., Mullineaux, P. M., and Davies, W. J.: Improving water use in crop production, Philos. T. Roy. Soc. B., 363, 639–658, 2008.

Neal, R. M.: Markov chain sampling methods for Dirichlet process mixture models, J. Comput. Graph. Stat., 9, 249–265, 2000.

Oncley, S. P., Foken, T., Vogt, R., Kohsiek, W., DeBruin, H., Bernhofer, C., Christen, A., Van Gorsel, E., Grantz, D., and Feigenwinter, C.: The energy balance experiment EBEX-2000. Part I: Overview and energy balance, Boundary Layer Meteorol., 123(1), 1–28, https://doi:10.1007/s10546-007-9161-1, 2007.

Ortega-Farias, S., Olioso, A., Fuentes, S., and Valdes, H.: Latent heat flux over a furrow-irrigated tomato crop using Penman–Monteith equation with a variable surface canopy resistance, Agric. Water Manage., 82, 421–432, 2006.

Ortega-Farias, S., Carrasco, M., Olioso, A., Acevedo, C., and Poblete, C.: Latent heat flux over a Cabernet Sauvignon vineyard using the Shuttleworth and Wallace model, Irrig. Sci., 25, 161–170, 2007.

Ortega-Farias, S., Poblete-Echeverria, C., and Brisson, N.: Parameterization of a two-layer model for estimating vineyard evapotranspiration using meteorological measurements, Agr. For. Meteorol., 150, 276–286, 2010.

Parlange, M. B., and Katul, G. G.: An advection-aridity evaporation model, Water Resour. Res., 28 (1), 127-132, 1992.

Poblete-Echeverria, C., and Ortega-Farias, S.: Estimation of actual evapotranspiration for a drip-irrigated Merlot vineyard using a three-source model, Irrig. Sci., 28, 65–78, 2009.

Priestley, C. H. B., and Taylor, R. J.: On the assessment of surface heat flux and evaporation using large-scale parameters, Mon. Weather Rev., 100 (2), 81-92, 1972.

Rana, G., Katerji, N., Ferrara, R.M., and Martinelli, N.: An operational model to estimate hourly and daily crop evapotranspiration in hilly terrain: validation on wheat and oat crops, Theory Appl. Climatol., 103, 413–426, 2011.

Sadegh, M., and Vrugt J. A.: Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM(ABC), Water Resour. Res., 50, 6767–6787, https://doi:10.1002/2014WR015386, 2014.

Samani, S., Ye, M., Zhang, F., Pei, Y. Z., Tang, G. P., Elshall, A. S., and Moghaddam, A. A.: Impacts of prior parameter distributions on bayesian evaluation of groundwater model complexity, Water Science & Engineering., 11(2), 89-100, https://doi.org/10.1016/j.wse.2018.06.001, 2018.

Schöniger, A., Wohling, T., Samaniego, L., and Nowak, W.: Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, Water Resour. Res., 50, 9484–9513, https://doi:10.1002/2014WR016062, 2014.

Schwarz, G.: Estimating the dimension of a model, Ann. Stat., 6(2), 461–464, https://doi:10.1214/aos/1176344136, 1978.

Shuttleworth, W. J., Gurney, R. J.: The theoretical relationship between foliage temperature and canopy resistance in sparse crops, Q. J. Roy. Meteorol. Soc., 116, 497–519, 1990.

Stannard, D. I.: Comparison of Penman-Monteith, Shuttleworth-Wallace, and modified Priestley-Taylor evapotranspiration models for wildland vegetation in semiarid rangeland, Water Resour. Res., 29 (5), 1379-1392, 1993.

Stull, R. B.: An introduction to boundary layer meteorology, Kluwer Academic Publ., 255pp, 1988.

Sumner, D. M., and Jacobs, J. M.: Utility of Penman–Monteith Priestley–Taylor reference evapotranspiration, and pan evaporation methods to estimate pasture evapotranspiration, J. Hydrol., 308 (1-4), 81-104, 2005.

Szilagyi, J., and Jozsa, J.: New findings about the complementary relationship based evaporation estimation methods, J. Hydrol., 354: 171–186, 2008.

Thomsen, J., Bohrer, G., Matheny, M. V., Ivanov, Y., He, L., Renninger, H., and Schäfer, K.: Contrasting hydraulic strategies during dry soil conditions in Quercus rubra and Acer rubrum in a sandy site in Michigan, Forests., 4(4), 1106–1120, 2013.

Tsvang, L., Fedorov, M., Kader, B., Zubkovskii, S., Foken, T., Richter, S., and Zeleny, Y.: Turbulent exchange over a surface with chessboardtype inhomogeneities, Boundary Layer Meteorol., 55(1–2), 141–160, 1991.

Vinukollu R, K., Wood, E. F., Ferguson, C. R., and Fisher, J. B.: Global estimates of evapotranspiration for climate studies using multi-sensor remote sensing data: evaluation of three process-based approaches, Remote Sens. Environ., 115(3), 801–823, 2011.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P. J., Hyman, M., and Robinson, B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resour. Res., 44, W00B09, https://doi:10.1029/2007WR006720, 2008.

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Higdon, D., Robinson, B. A., and Hyman, J. M.: Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, Int. J. Nonlinear Sci. Numer. Simul., 10(3), 273-290, 2009.

Webb, E. K., Pearman, G. I., and Leuning, R.: Correction of flux measurements for density effects due to heat and water-vapor transfer, Q. J. R. Meteorol. Soc., 106(447), 85–100, 1980.

Willmott, C. J.: On the validation of models, Phys. Geogr., 2, 184-194, 1981.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M. H.: Improving marginal likelihood estimaton for

Bayesian phylogenetic model selection, Syst. Biol., 60(2), 150-160, 2011.

Xu, C. Y., and Singh, V. P.: A review on monthly water balance models for water resources investigations, Water Resour. Manage., 12, 31-50, 1998.

Xu, Z. W., Liu, S. M., Li, X., Shi, S. J.,Wang, J. M., Zhu, Z. L., Xu, T. R., Wang, W. Z., and Ma, M. G.: Intercomparison of surface energy flux measurement systems used during the HiWATERUSOEXE, J. Geophys. Res., 118, 13140–13157, 2014.

Ye, M., Neuman, S. P., and Meyer, P. D.: Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, Water Resour. Res., 40, W05113, doi:10.1029/2003WR002557, 2004.

Ye, M., Meyer, P. D., and Neuman, S. P.: On model selection criteria in multimodel analysis, Water Resour. Res., 44, W03428, doi:10.1029/2008WR006803, 2008.

Zeng, X., Ye, M., Wu, J., Wang, D., & Zhu, X.: Improved nested sampling and surrogate‐enabled comparison with other marginal likelihood estimators, Water Resour. Res., 4, DOI: 10.1002/2017WR020782, 2018.

Zhang, K., Ma, J., Zhu, G., Ma, T., Han, T., and Feng, L. L.: Parameter sensitivity analysis and optimization for a satellite-based evapotranspiration model across multiple sites using Moderate Resolution Imaging Spectroradiometer and flux data, J. Geophys. Res. Atmos., 122, 230–245, https://doi:10.1002/2016JD025768, 2017.

Zhang, B., Kang, S., Li, F.,and Zhang, L.: Comparison of three evapotranspiration models to Bowen ratio-energy balance method for vineyard in an arid desert region of northwest China, Agr. Forest Meteorol., 148: 1629–1640, 2008.

Zhu, G. F., Su, Y. H., Li, X., Zhang, K., and Li, C. B.: Estimating actual evapotranspiration from an alpine grassland on Qinghai–Tibetan plateau using a two-source model and parameter uncertainty analysis by Bayesian approach, J. Hydrol., 476, 42–51, 2013.

Zhu, G. F., Li, X., Su, Y. H., Zhang, K., Bai, Y., Ma, J. Z., Li, C. B., Hu, X. L., and He, J. H.:

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Simultaneously assimilating multivariate data sets into the two-source evapotranspiration model by Bayesian approach: Application to spring maize in an arid region of northwestern China, Geosci. Model. Dev., 7(4), 1467–1482, 2014.

Zhang, X. Y., Liu, C. X., Hu, B. X., and Zhang, G. N.: Uncertainty analysis of multi-rate kinetics of uranium desorption from sediments, J. Contam. Hydrol., 156(1), 1-15, 2014.

Appendix A: List of symbols and physical characteristics in ET models

A	Available energy for the whole canopy (Wm^{-2})
A_s	Available energy ($W\ m^{-2}$)
R_n	Net radiation fluxes into the canopy ($W\ m^{-2}$)
R_{ns}	Net radiation flux into the substrate ($W\ m^{-2}$)
G	Soil heat flux ($W\ m^{-2}$)
λET	Sum of the latent heat flux from the crop (λT) and soil (λE) ($W\ m^{-2}$)
ET_c	Canopy transpiration ($W\ m^{-2}$)
ET_s	Soil evaporation ($W\ m^{-2}$)
C_c	Canopy resistance coefficient (dimensionless)
C_s	Soil surface resistance coefficient (dimensionless)
k_A	Extinction coefficient for available energy
LAI	Leaf area index
Q_{50}	Visible radiation flux ($W\ m^{-2}$)
D_{50}	Vapor pressure deficit (kPa)
D_a	Vapor pressure deficit at the reference height (D_a=e_s-e_a) (kPa)
Q_h	Flux density of visible radiation at the top of the canopy ($W\ m^{-2}$)
K_Q	Extinction coefficient
K_A	Extinction coefficient
f	Fraction of evaporation soil and total evaporation
λ	Latent heat of water evaporation ($MJ\ kg^{-1}$)
Δ	Slope of the saturated vapour pressure curve ($Pa\ K^{-1}$)
γ	Psychrometric constant ($kPa\ K^{-1}$)
ρ	Density of air ($kg\ m^{-3}$)
k	Karman constant (0.41)
e_s	Saturated vapor pressure (kPa)
e_a	Actual vapor pressure (kPa)
q^*	Saturation-specific humidity at air temperatur ($kg\ kg^{-1}$)
q	Specific humidity of the atmosphere ($kg\ kg^{-1}$)
b_1	Empirical constant ($s\ m^{-1}$)
b_2	Empirical constant ($s\ m^{-1}$)
β_1	empirical constant
β_2	empirical constant
θ	Soil water content ($m^3\ m^{-3}$)
θ_a	Critical water content at which plant stress starts ($m^3\ m^{-3}$)
θ_b	Water content at the wilting point ($m^3\ m^{-3}$)

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

θ_r	Residual soil water content (m^3 m^{-3})
θ_s	Saturated water content (m^3 m^{-3})
Θ	Relative water saturation
d	Zero plane displacement height (m)
z_m	Height of the wind speed and humidity measurements (3 m)
z_{0m}	Roughness length governing the transfer of momentum (m)
z_{0v}	Roughness length governing the transfer of water vapor (m)
h	Canopy height (m)
u_z	Wind speed at height z_m (m s^{-1})
g_a	Aerodynamic conductance (m s^{-1})
g_s	Surface conductance (m s^{-1})
g_{max}	Maximum stomatal conductance of leaves at the top of the canopy (m s^{-1})
g_s^c	Canopy conductance (m s^{-1})
r_a	Aerodynamic resistance (s m^{-1})
r_a^a	Aerodynamic resistance between canopy source height and a reference level (s m^{-1})
r_a^s	Aerodynamic resistance between the substrate and the canopy source height (s m^{-1})
r_a^c	Bulk boundary layer resistance of the vegetation element in the canopy (s m^{-1})
r_s^s	Surface resistance of the canopy (s m^{-1});
r_s^c	Bulk stomatal resistance of the canopy (s m^{-1})

Appendix B: Bayesian inference and the DREAM algorithm

The posterior probability distribution of the parameter is calculated by Bayes' theorem:

$$\pi(\theta \mid D, M) = \frac{\pi(\theta \mid M)\, p(D \mid \theta, M)}{p(D \mid M)} \qquad \text{(A1)}$$

where $\pi(\theta \mid M)$ represents the prior density of θ under model M; $p(D \mid \theta, M)$ is the joint likelihood of model M and its parameters θ; and

$$p\left(D \mid M\right) = \int p\left(D \mid \theta, M\right) p\left(\theta \mid M\right) d\theta \qquad \text{(A2)}$$

is the marginal likelihood, or Bayesian model evidence (BME).

The likelihood function, $p(D|\theta, M)$, used for parameter estimation, is specified according to the distributions of observation errors. Error $e(t)$ in each observation $D(t)$ at time t is expressed by

$$e(t) = D(t) - f(t) \tag{A3}$$

. Assuming $e(t)$ follows a Gaussian distribution with a zero mean, and the likelihood function can be expressed as

$$p(D|\theta) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{[e(t)]^2}{2\sigma^2}} \tag{A4}$$

where n is the number of observations and σ represents the error variances.

In this study, we used the DREAM algorithm (Vrugt et al., 2008, 2009) to explore the ET models' parameter space and to estimate BME. The DREAM sampling scheme is an adaptation of the global optimization algorithm of a shuffled complex evolution metropolis (SCEM-UA). This algorithm was descripted in more detail in Vrugt et al. (2008, 2009).

List of Tables

Table 1 Prior distributions and parameter limits for the PM, SW, PT-FC and AA models. The values are derived from the literature.

Table 2 Maximum Likelihood Estimates (MLEs), Mean Estimates, 95% High-Probability Intervals (Lower Limit, Upper Limit).

Table 3 Slope and coefficient of determination (R^2) of regression between measured and modeled half-hourly evapotranspiration values, and statistics of root mean square error (RMSE), mean bias error (MBE), index of agreement (IA), model efficiency (EF) and Logarithm of BME for the four ET models.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences
Open Access
Discussions

List of Figures

Table 1 Prior distributions and parameter limits for the PM, SW, PT-FC and AA models. The values are derived from the literature.

Parameter	Description	Prior range	PM	Prior for SW	Prior for PT and AA	References

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-430
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 22 October 2018

Hydrology and Earth System Sciences Discussions
Open Access

		Lower	upper	Lower	upper	Lower	upper	
g_{max} (mm s^{-1})	maximum stomatal conductance	0	50	0	50			Kelliher et al. (1995)
Q_{50} (W m^{-2})	visible radiation flux	10	50	10	50			Leuning et al. (2008)
D_{50} (kPa)	vapor pressure deficit	0.5	3	0.5	3			Leuning et al. (2008)
K_Q	extinction coefficient	0	1	0	1			Leuning et al. (2008)
K_A	extinction coefficient	0	1	0	1			Leuning et al. (2008)
b_1 (s m^{-1})	empirical constant			4.5	11.3			Sellers et al. (1992)
b_2 (s m^{-1})	empirical constant			0	8			Sellers et al. (1992)
β_1	empirical constant					0.5	1.5	Flint et al. (1991);
B_2	empirical constant					0.1	10	Barton. (1979)

Table 2 Maximum Likelihood Estimates (MLEs), Mean Estimates, 95% High-Probability Intervals (Lower Limit, Upper Limit).

Parameter	Posterior for PM			Posterior for SW			Posterior for PT and AA		
	MLE	Mean	CI	MLE	Mean	CI	MLE	Mean	CI
g_{max} (mm s^{-1})	0.04	0.04	(0.03, 0.04)	0.01	0.01	(0.005, 0.012)			
Q_{50} (W m^{-2})	49.96	48.52	(39.73, 49.74)	47.49	40.32	(11.02, 48.99)			
D_{50} (kPa)	3.00	2.87	(1.92, 2.97)	2.98	2.88	(2.26, 2.98)			
K_Q	1.00	0.99	(0.911, 0.998)	0.99	0.88	(0.06, 0.98)			
K_A	1.00	0.98	(0.822, 0.995)	0.12	0.12	(0.074, 0.184)			
b_1 (s m^{-1})				4.51	4.57	(4.52, 4.96)			
b_2 (s m^{-1})				0.39	0.57	(0.07, 1.38)			

β_1	1.1[a] 1.5[b]	1.098[a] 1.499[b]	(1.06, 1.16)[a] (1.492, 1.499)[b]
β_2	10.00[a] 10.00[b]	9.75[a] 9.94[b]	(7.97, 9.95)[a] (9.44, 9.99)[b]

[a] PT-FC model; [b] AA model.

Table 3 Slope and coefficient of determination (R^2) of regression between measured and modeled half-hourly evapotranspiration values, and statistics of root mean square error (RMSE), mean bias error (MBE), index of agreement (IA), model efficiency (EF) and Logarithm of BME for the four ET models.

Model	Slope	R^2	RMSE	MBE	IA	EF	BME
PM	1.01	0.76	85.38	-9.52	0.93	0.74	-6300.5
SW	1.05	0.82	76.34	-19.07	0.95	0.79	-6025.1
PT-FC	0.91	0.75	94.39	25.42	0.92	0.68	-6366.8
AA	0.92	0.75	95.09	23.29	0.92	0.67	-6390.3

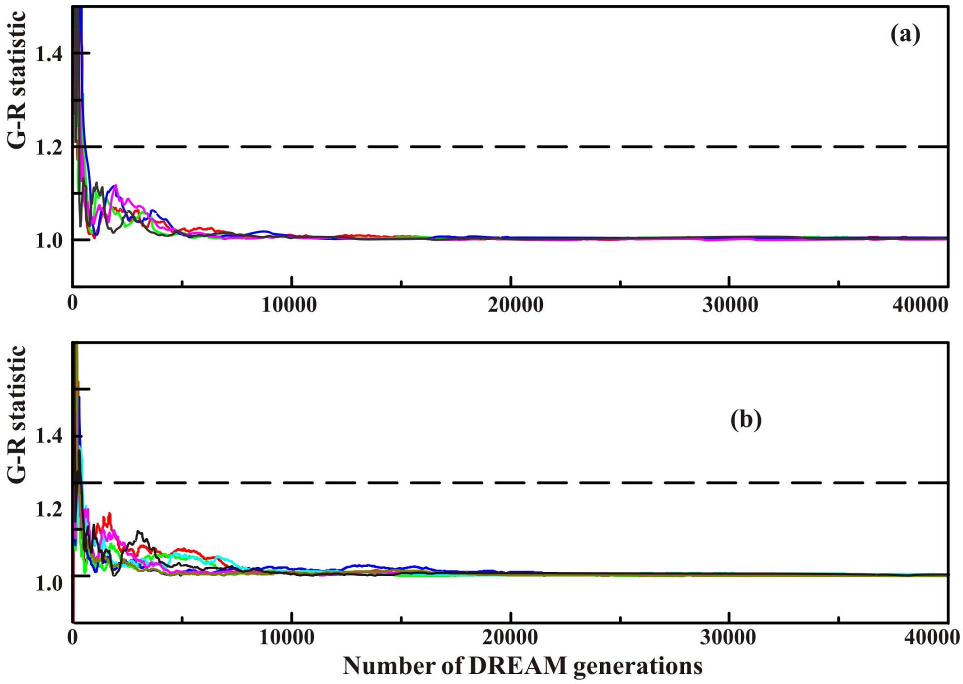


Figure 1 Trace plots of the G-R statistic of Gelman and Rubin (Gelman and Rubin, 1992) using DREAM for the PM model (a) and (b) the SW model. Different parameters are coded with different colors. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution.

Hydrology and Earth System Sciences
Open Access
Discussions

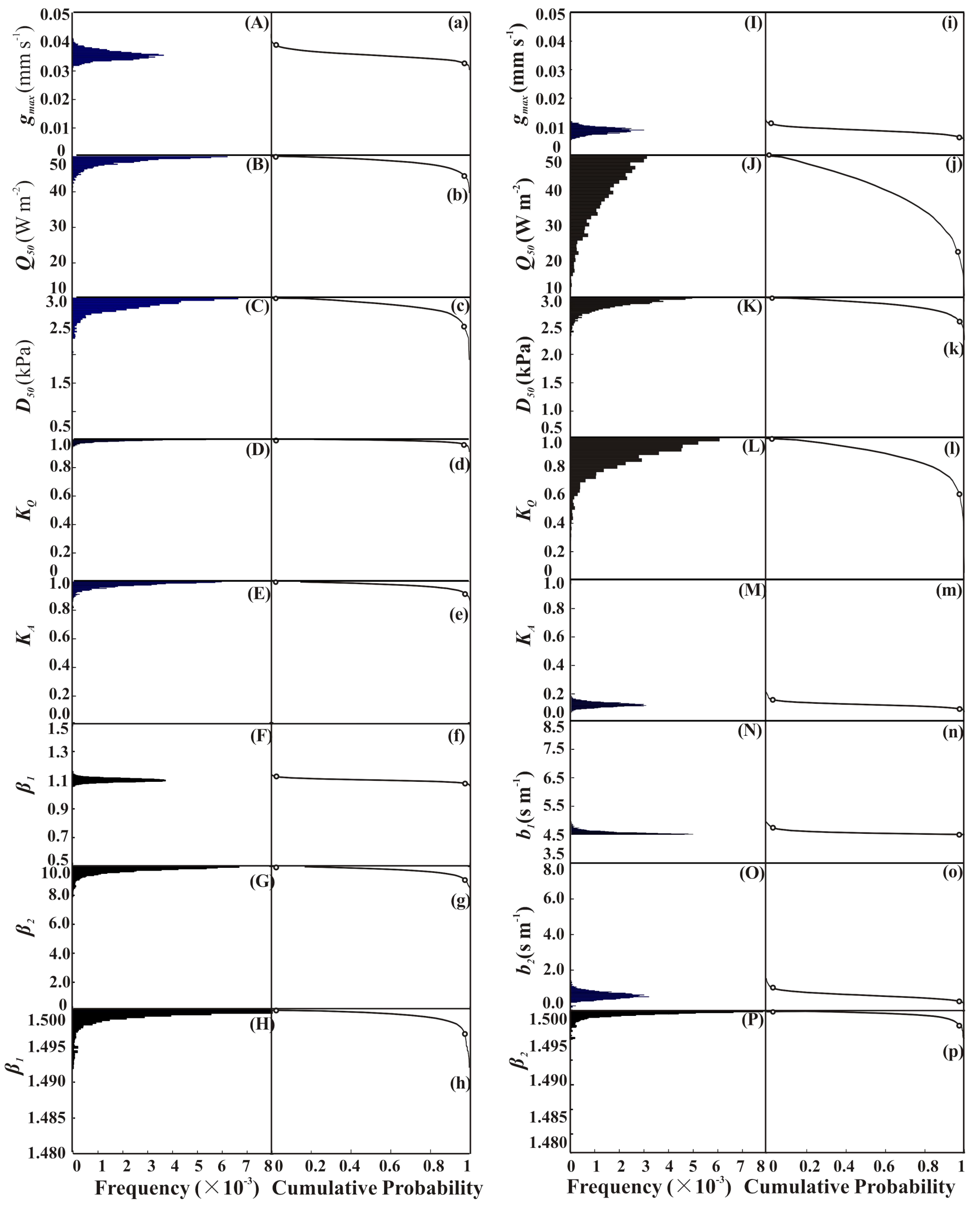
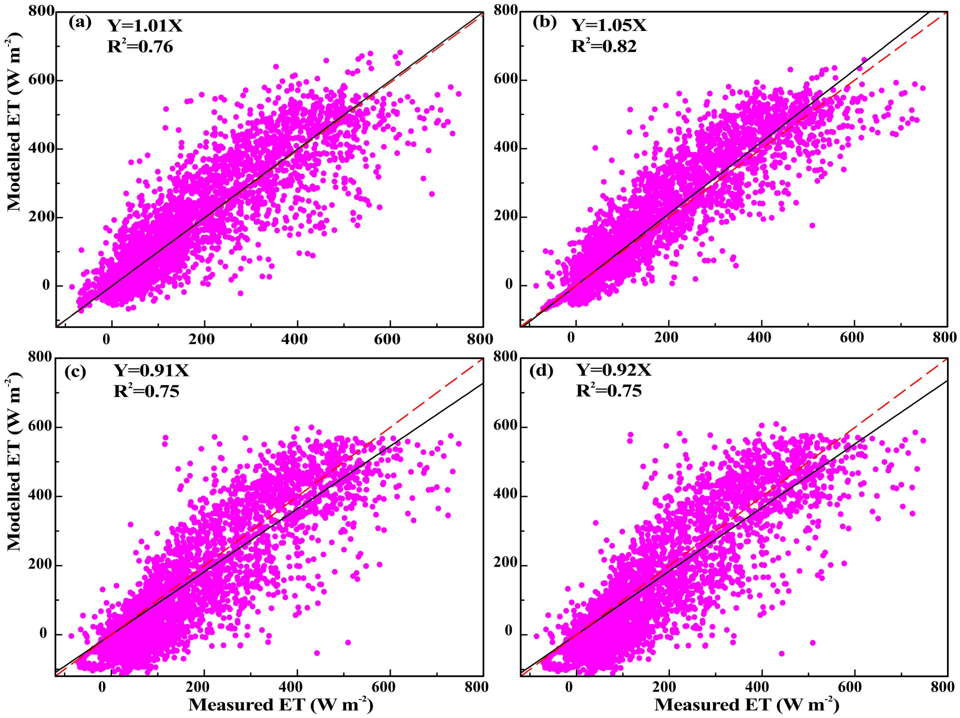
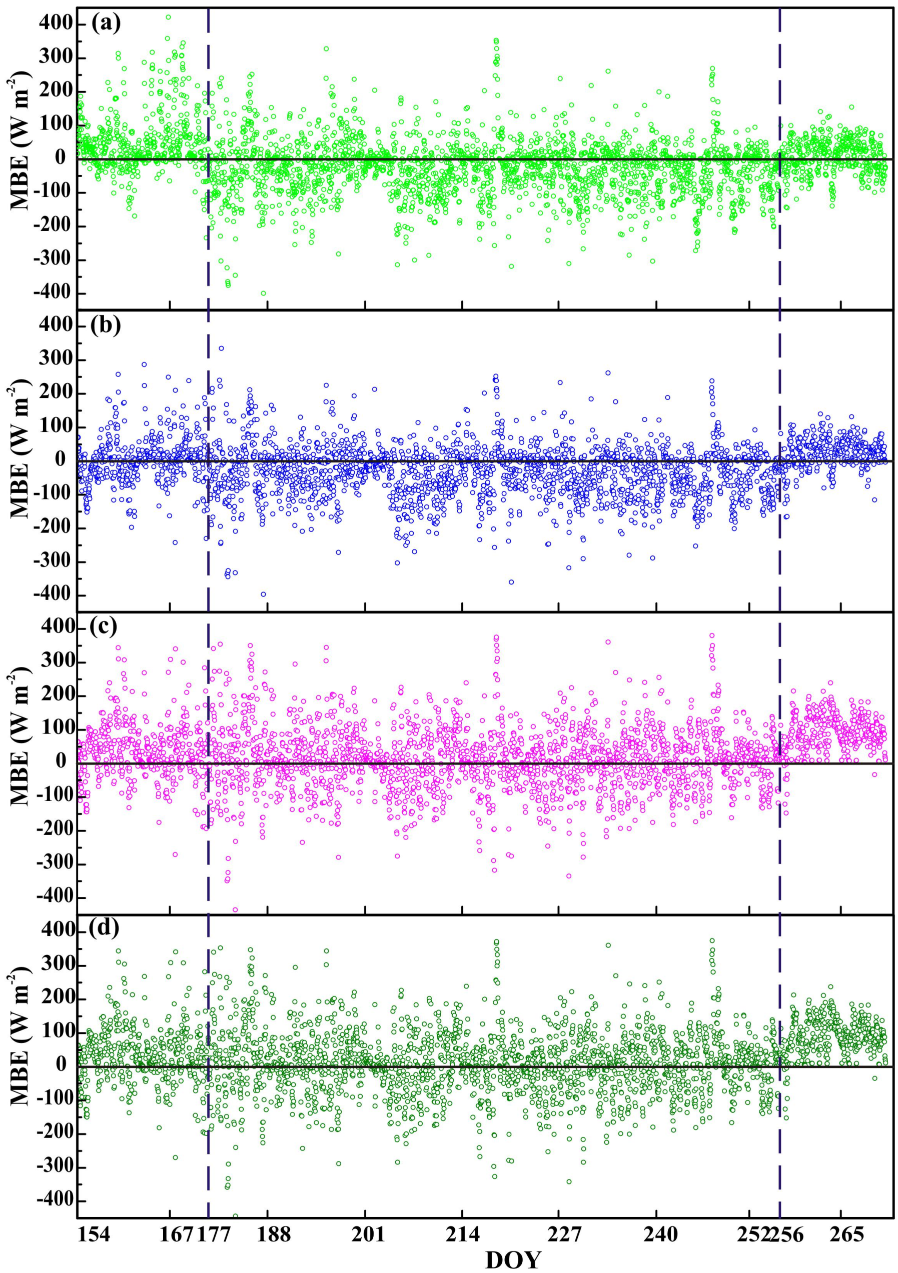


Figure 2 Uppercase of 2A-2E, 2I-2O, 2F-2G, and 2H and 2P show histograms, and corresponding lowercases of 2a-2e, 2i-2o, 2f-2g, and 2h and 2p show CDFs for the PM, SW, PT-FC and AA models, respectively. Thee histograms and the CDFs are constructed from the one chain and 40000 generations simulated using DREAM. The *y* axes represent the prespecified limits of the parameters.



850

851 **Figure 3.** Regression between measured and modeled half-hourly evapotranspiration values produced by
852 different models: (a) PM, (b) SW, (c) PT-FC and (d) AA. The regressions are: Y = 0.99X (R^2 = 0.76), Y =
853 1.05X (R^2 = 0. 82), Y = 0.91X (R^2 = 0.75), and Y = 0.92X (R^2 = 0.75) for the PM, SW, PT-FC and AA
854 models, respectively.

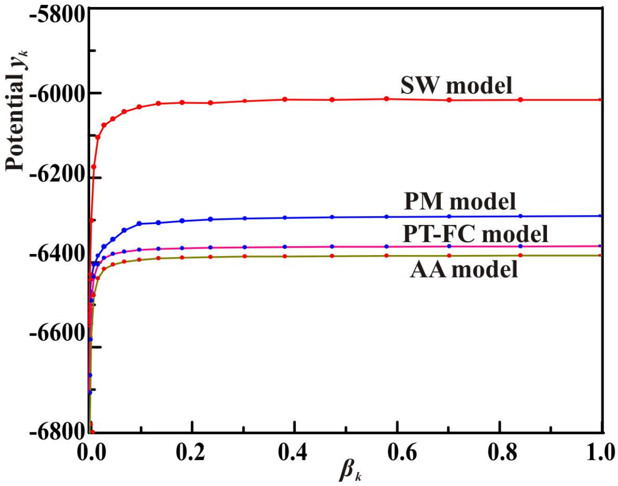


855 Fig

856 **ure 4.** Mean bias error (MBE) of predicted and observed ET (W m^{-2}) values for (a) PM, (b) SW, (c) PT-FC

857 and (d) AA models from DOY 154 to DOY 270. Parameters used for prediction are estimated by DREAM

858 with the dataset for the calibration period from DOY 154 to DOY 202.

859 

860 **Figure 5.** Variation of the mean posterior expectation of the potential y_k (equation (36)) with β_k (power
861 coefficient in equation (33)) for the PM, SW, PT-FC and AA models.

862