Manuscript Number: hess-2018-430

Title: Bayesian performance evaluation of evapotranspiration models：a case study based on eddy covariance system of a maize field in northwestern China

Corresponding Author: Xiaoying Zhang

Authors: GuoxiaoWei, Xiaoying Zhang, Ming Ye , Ning Yue, Fei Kan

Dear Editor,

On behalf of my co-authors, we thank you very much for giving us an opportunity to revise our manuscript. We are grateful to the editors and reviewers for their positive and constructive comments and suggestions on our manuscript (hess-2018-430) entitled “Bayesian performance evaluation of evapotranspiration models for an arid region in northwestern China”.

We have studied reviewer’s comments carefully, and revised the manuscript thoroughly to address the comments. The revision is marked in red in the revised manuscript. We have tried our best to revise our manuscript according to the comments. Attached please find the revised version, which we would like to submit for your kind consideration.

We would like to express our great appreciation to you and reviewers for comments on our paper. Looking forward to hearing from you.

Thank you and best regards.

Yours sincerely,

Xiaoying Zhang

Corresponding author:

Name Xiaoying Zhang

E-mail: xiaoyingzh@jlu.edu.cn

List of Responses

Dear Editors and Reviewers:

Thank you for your letter and for the reviewers' comments concerning our manuscript entitled "Bayesian performance evaluation of evapotranspiration models for an arid region in northwestern China" (hess-2018-430). These comments are all valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our researches. We have studied comments carefully and have made corrections which we hope meet with approval. Revised portion are marked in red in the paper. The main corrections in the paper and the responds to the reviewer's comments are followed.

Thank you and best regards.

Yours sincerely,

Xiaoying Zhang
Corresponding author:
Name: Xiaoying Zhang
E-mail: xiaoyingzh@jlu.edu.cn

Responds to the comments:

Referee #1:

General comments:

1. Comment: *For the objective (1), what is the purpose of selecting the best model using BME, to improve model prediction? If the purpose is to improve model prediction, did the authors try Bayesian model averaging? Based on the results, some models are underestimate, some models are overestimate, it is possible that model averaging could give a better prediction performance.*

Response: We believe this comment is very important for considerably improving our manuscript. Our original idea was to identify which model is optimal for ET prediction, and then to improve the model with the Bayesian model averaging. However, our result showed that model SW obtained a weight of 100%. This means that BME assigned a weight of 100% to the competing model and the weights of 0% to other three alternative models, and thus, the Bayesian model averaging prediction is also the SW prediction. Other studies on hydrological model selection have yielded similar results in that one model obtained an weight of close to 100% (e.g., Meyer et al., 2007; Lu et al., 2013; Schöniger et al., 2007). Therefore, Bayesian model averaging was not used in our study.

2. Comment: *For the objective (2), theoretically we know these statistics only measure model fit without considering model complexity, so they are not as robust as BME. And we know these statistics can be efficiently calculated, so there is no need for testing. Please justify the objective (2).*

Response: We have changed the original objective (1) and (2) to "(1) to calibrate ET model parameters using the diffeRential evolution adaptive metropolis (DREAM) algorithm; (2) to identify which parameters had a greater impact on the model performance and to explain why the selected optimal model performed best". These changes can be seen at P5, L120-121.

3. Comment: *objective (3) is very meaningful. I would like to see more analysis on the model-data mismatch to improve model development and model performance.*

Response: Considering the Reviewer's suggestion, we added some sentences in 4.1 "Parameter uncertainty analysis", which can be seen at P18-P19, L439-453, and in 4.3 "Analysis of model-data mismatch", which can be seen at P24, L597-601.

4. Comment: *In several places of the manuscript, the logic is not very clear. The English writing needs improvement.*

Response: It is true as you suggested that our manuscript needs the revision of English sentence. After the revision according to the reviewer's comments, the manuscript have been edited by the professional translation services.

Specific comments:

1. Comment: *Abstract, I think including some insights obtained from the numerical experiments in the abstract would attract more audience and make this work more meaningful.*

Response: Considering the Reviewer's suggestion, we added and revised the content in abstract. These changes can be seen at P1, L19-23.

2. Comment: *2. Line 35-36, the SW model performs best in this study area, but may not be the best in other areas. For example, in Li et al., (2013)'s study, PM performed better than SW in estimation of maize. Please justify the statement that SW should be the first choice for evaluating ET of spring maize in arid desert oasis areas.*

Response: We have changed this statement to "The mismatch analysis indicated that explicit treatment of energy imbalance and energy interaction will be the primary way to further improve ET model performance." The change can be seen at P2, L38-39.

3. Comment: *Line 93-94, BME can be used to compare and select the best-performing model. This is well-known and not a hypothesis that needs to be determined. What do the authors mean by saying "an unbiased view"?*

Response: We have deleted this sentence at P4, L111.

4. Comment: *Line 95-97, the first part of the sentence says Bayesian applications have focused on comparison of alternative models, but the second part of the sentence says that little attention has been given to the Bayesian model comparison. The sentence is self-contradictory. Please clarify.*

Response: Our original intention is to say that Bayesian applications have focused calibration of individual models but the model comparison is still conducted using traditional statistical criteria. Our expression was not very clear. We changed this statement as "Most applications of Bayesian methods have focused on the calibration of individual models, while the comparison of alternative models continues to be performed using traditional error metrics." This changes can be seen at P4, L107-108.

5. Comment: *Line 277, for each chain? I thought you total have 40,000 samples from all chains. In addition, Line 848, from one chain? Please clarify.*

Response: This referred the 40,000 samples from one chain. Total have 40,000 multiplied by *N* (chain number) samples from all chains. Please see P13, L298-299; P32, L874; P35, L905.

6. Comment: *Line 280-282, based on Figure 1, DREAM needs far less than 8000 generations to make the GR statistic smaller than 1.2. Also, based on Figure 1's x-axis scale, it is hard to tell "obviously" the chain converged after about 620 and 450 generations.*

Response: Thanks for the comment. We changed the sentence "DREAM needs far less than 8000 generations to make the GR statistic smaller than 1.2 for the both models " to "The algorithm needs about 8,000 generations to make the G-R statistic close to 1.0 for the both models." In addition, we deleted statement "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 change can be seen at P13-14, L302-304.

7. Comment: *Figure 1. In Figure 1(b) the position of the dash line is not at 1.2. The position of the label (b) is not aligned well with the label (a).*

Response: Thanks for the comment. We have redrawn the Figure 1. This can be seen Figure 1.

8. Comment: Figure 2. If the authors cannot get more information from the CDFs than the histograms, I suggest deleting the CDFs which make Figure 2 busy and confusing. Also, I would like to see more discussion about Figure 2; what insights the authors can obtain from these plots?

Response: We have redrawn the Figure 2. This can be seen Figure 2.

The main insights summarized as following: g_s and modeled ET in PM model are relatively insensitive to Q_{50}, D_{50} and K_q. Hence, these parameters could not be well constrained. The calculation of g_s^c in SW model is the same as in PM model, and thus, g_s^c and modeled ET for SW model are also insensitive to parameters of Q_{50}, D_{50}, K_q. Therefore, these three parameters were also not be well constrained in SW model. In addition, for edge-hitting parameters, their uncertainties may be also the outcome of model biases or EC-measured ET data errors. Although the ecophysiological parameter g_{max} is a variable in the g_s^c equation in both PM and SW models, but this parameter is sensitive to g_s^c and has large influences on the evaluated ET. Its effect is relatively independent compared to the other meteorological parameters in the models, and therefore this parameter was well specified in SW model. The parameter K_a is insensitive to g_s and modeled ET. In contrast, K_a is contained in equation of net radiation flux into the substrate (Eq.12) in SW model. From the above analysis, we could see that K_a not only involved the distribution of energy between the canopy and the soil surface but also the energy imbalance. Therefore, parameter K_a has a great influence on the performance of the SW model.

The revisions and changes can be seen P14, L314-321; P14, L326-330 and P17, L397-412.

9. Comment: *Line294-297, I found the discussion of the figure 2 is confusing. I think, the figure 2 says the histograms tend to concentrate in the upper bounds, not the lower bounds. Also, the authors should increase the upper limits of these parameters not decrease, because the histograms are concentrated in the upper bounds.*

Response: Thanks for the point. We have revised this sentence. The change can be seen at P14, L315-321.

11. Comment: *Line 355-356, what do the author mean by saying "to sample groups of variable in turn"?*

Response: The sentence should be "to sample one or groups of variable in turn". We have removed some sentence and this can be seen P17, L388-393.

12. Technical corrections: *Line 29, obstained –> obtained. Line 92, beed –> been*

Response: We have corrected the words already.

Once again, thank you very much for your comments and suggestions.

Referee #2:

Comments:

1. Comment: *The grammar of this paper needs some improvements, some grammar errors and ambiguous sentences can be found.*

Response: It is really true as you suggested that our manuscript needs a language improvement. After the revision according to the reviewer's comments, the manuscript have been edited by the professional translation services.

2. Comment: *The universality of this study and its conclusions need to be clarified since the study area and methodology are both very spatial and temporal specific. PS, are the conclusions valid under other conditions or not?*

Response: The ET models and BME model selection can be applied to other conditions as long as the required data can be obtained. Although there are many studies on ET model evaluation, their conclusions about model ranking are all based on traditional error metrics. Just as you said, the conclusion about whether SW model is optimal selected by BME method under other conditions still needs further confirmation. We have added relevant contents at P24, L594-596.

3. Comment: *Following the last comment, is it possible to provide results for other study areas or using other time scales? This will provide strong evidences to support the conclusions.*

Response: It is really true that providing results for other study areas or using other time scales would be very useful for providing strong evidences to support the conclusions. We've been looking for reliable data from other study area or from other crops for BME model selection to confirm whether the SW model is the optimal model under other conditions. However, it is difficult to obtain the required data by ET models, especially the soil water contents. So far, we haven't got the requited data yet. And thus, Thanks for the comment that we are not able to provide results for other study areas or using other time scales for BME model selection by now.

4. Comment: *I am not sure I can agree with some conclusions, for example, the one in lines 531-532, the authors suggest prioritizing BME over other measurements, but BME can also provide inaccurate results.*

Response: We think this is true, and we deleted statement "and that BME should be used instead", and reorganized the original sentences. Please see P25, L609-615.

Once again, thank you very much for your comments and suggestions.

Referee #3:

General Comments:

1. Comment: *Language issues should be fully checked throughout the entire text before publication in HESS.*

Response: It is really true as you suggested that our manuscript needs the revision of English sentence. After the revision according to the reviewer's comments, the manuscript have been edited by the professional translation services.

2. Comment: *Novelty of the paper should be better emphasized rather than "BME has not been used for evaluating the ET models".*

Response: Thanks for the comment. We have changed the original statement to "Currently, ET model selection and comparison have been still conducted using traditional error metrics. It is known that error metrics are not adequate to provide reasonable result of model ranking for disregarding model complexity (Marshall et al., 2005; Samani et al., 2018). The focus of this study is to use a Byesian approach to evaluate the performance of the PM, SW, PT-FC, and AA models, which is a novelty contribution of this study." These changes can be seen at P3, L81-85.

3. Comment: *Model complexity for each model should be better described. For example, authors can directly introduce number of parameters with uncertainties in their experiment?*

Response: Considering the comments, we have described the number of parameters of each model at P13, L289-291 and added a sentences "The results illustrate that with the addition of parameters, the model complexity and the model performance are both increased." at P16, L382-383.

Specific comments:

1. Comment: *The Abstract is out of organization. It seems to me that you never mention the model complexity but always write "underestimation" or "overestimate" to explain why the SW is the best one.*

Response: We have reorganized the Abstract. These can be seen P1-P2, L19-39.We also have added sentence "Although the SW model with seven parameters is sophisticated, it's good fitting to observations can counterbalance its higher complexity." These can be seen P2, L35-36.

2. Comment: *Lines 25-27: unclear, please rephase this sentence.*

Response: Considering the comments, we have reorganized the abstract again, and changed the original sentence to "The parameters in each model were first calibrated using DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm, and then were analyzed to identify their impacts on the model performance. The Bayesian model evidence (BME) approach, was further adopted to select the optimal model by incorporating the mathematically rigorous thermodynamic integration algorithm." These modifications can be seen at P1, L19-23.

3. Comment: *It is unclear for me why 'SW' is best one from the abstract.*

Response: Considering the comments, we have added the sentences "Our results revealed that the extinction coefficient was the most significant parameter in the ET models. It was not merely partitioning the total available energy into the canopy and surface, but also including the energy imbalance correction. The extinction coefficient is well constrained in the SW model and poorly constrained in the PM model, but not considered in PT-FC and AA models" to explain why SW is the best one in abstract. Please see P2, L31-35.

4. Comment: *Line37: please check the symbol.*

Response: We have checked the symbol at P2, L40.

5. Comment: Simulate ET or estimate ET? Please be very sure of this word.

Response: We have changed some "estimate ET" to "simulate ET".

6. Comment: *Line 41: add a reference.*

Response: Thanks for the comments. We have added corresponding reference "(Brutsaert, 2005)". Please see at P2, L45.

7. Comment: *Lines 55-56: unclear, please rephase this sentence.*

Response: We rephased this sentence as "These ET models are generally complex, because of for the coupling of the land surface and atmospheric processes, and high-dimensional with a large number of parameters". Please see at P3, L59-60.

8. Comment: *Lines 62-63: 'These quantitative criteria' refer to what?*

Response: We have reorganized this paragraph, and deleted the original statement. Please see at P3, L67-68.

9. Comment: *Line 70: performances.*

Response: Thanks. We have reorganized this paragraph and corrected the statement. Please see at P3, L75.

10. Comment: *Line 71: remove 'the' from 'the SW model'*

Response: We have removed 'the' from 'the SW model'.

11. Comment: *Lines 71-72: please rephase this sentence.*

Response: We have changed the sentence "Ershadi et al. (2014) evaluated the surface energy balance system (SEBS), PM, PT-JPL (a modified Priestley-Taylor model), and AA models." This can be seen at P3, L76-77.

12. Comment: *Line 73: should be model ranking? Please check the terminology.*

Response: Thanks. We have corrected this mistake at P3, L69.

13. Comment: *Lines 75-76: unclear, significant variability of model performances?*

Response: Considering the comments, we have changed "significant" to "considerable". This can be seen at P3, L80.

14. Comment: *Lines 92-93: been?*

Response: We have corrected the sentence.

15. Comment: *Lines 102-103: add a reference*

Response: We have added the reference "(Vrugt et al., 2008, 2009)" at P5, L119-120.

Once again, thank you very much for your comments and suggestions.

Bayesian performance evaluation of evapotranspiration models: a case study ~~for an arid region~~based on eddy covariance system of a maize field in northwestern China

Guoxiao Wei [1, 2 *], Xiaoying Zhang [3, *], Ming Ye [4], 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] Construct Engineering College, Jilin University, China, 130400
[4] Department of Earth, Ocean, and Atmospheric Science, Florida State University, USA, 32306
* Corresponding author: xiaoyingzh@jlu.edu.cn.

Abstract

Evapotranspiration (ET) is a major component of the land surface process involved in energy fluxes and energy balance, especially in the hydrological cycle of agricultural ecosystems. While many models have been developed as powerful tools to estimate ET, there ~~has been~~is no agreement on which model ~~has the~~ best describing the loss of water to the atmosphere. In this study, we present a solid study to evaluate four widely used ET models and their parameter contributions ~~(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 four tested models are ~~the~~ Shuttleworth Wallace (SW) model, Penman-Monteith (PM) model, Priestley-Taylor and Flint-Childs (PT-FC) model, and Advection-Aridity (AA) model. The parameters in each model were first calibrated using DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm, and then were analyzed to identify their impacts on the model performance. The Bayesian model evidence (BME) approach, was further adopted to select the optimal model by incorporating the mathematically rigorous thermodynamic integration algorithm.~~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 obstained 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.~~ Our results revealed that the extinction coefficient was the most significant parameter in the ET models. It was not merely partitioning the total available energy into the canopy and surface, but also including the energy imbalance correction. The extinction coefficient is well constrained in the SW model and poorly constrained in the PM model, but not considered in PT-FC and AA models. ~~icient is well constrained in the SW model and poorly constrained in the PM model, but not considered in PT-FC and AA models.~~ This is the main reason that the SW model outperforming the other models. Although the SW model with seven parameters is sophisticated, it's good fitting to observations can counterbalance its higher

complexity. In addition, the discrepancies between observations and model simulations were evaluated using traditional error metrics. The mismatch analysis indicated that explicit treatment of energy imbalance and energy interaction will be the primary way to further improve ET model performance.

Keywords: Bayesian analysis; ET models; Eddy covariance; ~~Penman–Monteith; Shuttleworth-Wallace;~~ Model perfornace; Extinction coefficient~~; Maize~~

1. Introduction

Surface energy fluxes are an important component of Earth's global energy budget and a primary 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 (Brutsaert, 2005). In agricultural ecosystems, more than 90% of the total water losses are due to ET (~~Brutsaert~~Morison et al., ~~2005~~2008). Therefore, robust ET estimation is crucial to a wide range of problems in hydrology~~ (Xu and Singh, 1998)~~, ecology, and global climate change (Xu and Singh, 1998~~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~~ between 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 only require a small ~~amount~~number of ground-based measurements ~~for setting~~to set up the models (Ershadi et al., 2014).

These ET models are generally complex, because of ~~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~~ Comparing the performance of competing models ~~and how~~and ~~to~~ evaluating ~~e~~and understanding the ~~mismatch~~ discrepancies between simulations of the –model~~ simulation~~s and corresponding observed surface-atmosphere water flux are remain challenging problems (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~~ to ~~evaluating~~ evaluate the ~~model~~ performance of ET models (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~~ simulations of the PM, SW and adjusted SW models under film-mulching conditions of maize growth 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~~ adjusted SW models underestimated the daily ET by 6% and 2%, respectively~~, during the entire experimental period of 116~~

~~days~~. Therefore, the performances of PM and adjusted SW models are better than that of the SW model in their case study. Ershadi et al. (2014) evaluated the surface energy balance system (SEBS), PM, PT-JPL (a modified Priestley–Taylor model~~, similar to the PT-FC~~) and AA models. Based on the average value of EF and RMSE, the model ranking from ~~the~~ worst to ~~the~~ best was AA, PM, SEBS, and PT-JPL. Ershadi et al. (2015) also compared the ~~evaluated model~~ response of the models to ~~the~~ different formulations of aerodynamic and surface resistances ~~against~~ with global FLUXNET data. Their results showed ~~significant~~ considerable variability in model performance among and within biome types. Currently, ET model selection and comparison have been still conducted using traditional error metrics. It is known that error metrics are not adequate to provide reasonable result of model ranking for disregarding model complexity (Marshall et al., 2005; Samani et al., 2018). The focus of this study is to use a Byesian approach to evaluate the performance of the PM, SW, PT-FC, and AA models, which is a novelty contribution of this study.

In ET models, the land surface energy system is governed by presumably infinite-dimensional physics. However, considering the ET models as finite-dimensional can be more precisely by covering all relevant relations. Therefore, employing consistent criteria for model selection might be justified when the aim is to better understand the processes involved (Höge et al., 2018). When using consistent model selection, ~~The~~ Bayesian model evidence (BME), also known as marginal likelihood, measures the average fit of ~~a~~ model simulations to their corresponding observations ~~to the data~~ over a model's prior parameter space. This feature enables BME to consider model complexity (in terms of number of model parameters) for model performance evaluation. 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, the Bayesian information criterion (BIC; ~~Kashyap~~Schwarz, ~~1982~~1978) ~~or~~ and the Kashyap information criterion (KIC; ~~Schwarz 1978~~Kashyap, 1982) ~~were~~ have been used to approximate BME ~~for reducing~~by using maximum likelihood theories to reduce computational cost of evaluating BME (Ye et al., 2004). However, these approximations have theoretical and computational limitations (Ye et al., 2008; Xie et al.~~,~~ 2011; Schöniger et al., 2014), and a numerical evaluation (not a likelihood~~n~~ 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), in order 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~~numerically accurate than the generally used harmonic mean method (Xie et al., 2011).

Most applications of Bayesian methods have focused on the calibration of individual models, while the comparison of alternative models continues to be performed using traditional error metrics. ~~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.~~ More generally, Bayesian approaches to ~~M~~model calibration, comparison, and analysis ~~underlying the Bayesian paradigm has~~ have been used ~~much~~ far less used in the evaluation of ET models than in other areas of environmental science. In this study, the Bayesian approach ~~was~~ is 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 objectives of the study are as follows: (1) ~~The~~ to calibrate ET model parameters ~~were calibrated~~ using the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al., 2008, 2009)~~.~~; ~~The objectives of the study are as follows:~~ (~~1~~2) to identify which parameters had a greater impact on the model performance and to explain why the selected optimal model performed best; ~~compare the four models and select the best one using BME;~~ (~~2~~3) to evaluate the performance of the models using traditional error metrics and BME; and ~~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~~4) to analyze discrepancies between model simulations and observation data in order to ~~model-data mismatch for~~ better understand~~ing~~ model performance and identify ways to improve these models. ~~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 of maize growth was conducted at Daman Superstation, located in Zhangye City, Gansu province, 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~~region that consumes large amount of water ~~consumption~~ for both domestic and agricultural uses. The annual average precipitation and temperature are 125 mm and 7.2 ℃ (1960–2000), respectively. The annual accumulated temperature (>10 ℃) is 3,234 ℃, 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~~ agricultural region, and the major source of water ~~resources are~~ is ~~the~~ snowmelt from the Qilian Mountains. ~~The maize~~ Maize and spring wheat are the principal crops grown in the region.~~,~~ 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 in the study area.

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~~a height of 4.5 m. Maize is the main crop in the study region, and thus covers ~~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~~methods, 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. Leaf Area Index (LAI) was measured approximately every 10 days during the growing season.

2.3. Model description

In this section, we summarize the mathematical definitions forming the basis of each of the four models. Appendix A contains a summary of the names and physical meanings of the model parameters.

2.3.1 Penman-Monteith (PM) model

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

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

where $\varepsilon=\Delta/\gamma$; and A is defined to be $A=R_{\mathrm{n}}-G$.

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

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

(~~3~~2)

where 1-τ and τ ~~is~~ are the fraction of the total available energy absorbed by the canopy and by the soil, and τ = exp (- $K_a LAI$), ~~;~~ and g_i and $g_{\mathrm{s}}^{\mathrm{c}}$ ~~is~~ are defined ~~as~~in equations (3) and (4), respectively (Monteith, 1965):

$$g_i = \frac{A}{\left(\rho C_p/\gamma\right) D_a}$$ (~~4~~3)

~~(Monteith, 1965); g_s^c is expressed as:~~

$$g_s^c = \frac{g_{max}}{K_q} In\left[\frac{Q_h + Q_{50}}{Q_h \exp(-K_q \mathrm{LAI}) + Q_{50}}\right]\left[\frac{1}{1 + D_a/D_{50}}\right] f(\theta)$$ (~~5~~4)

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

$$f(\theta) = \begin{cases} 1 & \theta > \theta_a \\ \dfrac{\theta - \theta_b}{\theta_a - \theta_b} & \theta_b < \theta < \theta_a \\ 0 & \theta < \theta_b \end{cases}$$ (~~6~~5)

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

$$g_a = \frac{k^2 u_m}{\ln\left[(z_m - d)/z_{0m}\right]\ln\left[(z_m - d)/z_{0v}\right]}$$ (~~7~~6)

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:

$$\lambda ET = \lambda E + \lambda T = C_s ET_s + C_c ET_c \tag{7}$$

$$ET_s = \frac{\Delta A + \left\{\rho C_p (e_s - e_a) - \Delta r_a^s (A - A_s)\right\}/\left(r_a^a + r_a^s\right)}{\Delta + \gamma\left\{1 + r_s^s/\left(r_a^a + r_a^s\right)\right\}} \tag{8}$$

$$ET_c = \frac{\Delta A + \left\{\rho C_p (e_s - e_a) - \Delta r_a^c A_s\right\}/\left(r_a^a + r_a^c\right)}{\Delta + \gamma\left\{1 + r_s^c/\left(r_a^a + r_a^c\right)\right\}} \tag{9}$$

where the available energy input above the soil surface is defined as $A_s = R_{ns} - G$.

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

$$R_{ns} = R_n \exp(-K_A \mathrm{LAI}) \tag{10}$$

The coefficients C_s and C_c are obtained as follows:

$$C_{\mathrm{s}} = \left\{1 + R_{\mathrm{s}} R_{\mathrm{a}} / R_{\mathrm{c}} \left(R_{\mathrm{s}} + R_{\mathrm{a}}\right)\right\}^{-1} \tag{11}$$

$$C_{\mathrm{c}} = \left\{1 + R_{\mathrm{c}} R_{\mathrm{a}} / R_{\mathrm{s}} \left(R_{\mathrm{c}} + R_{\mathrm{a}}\right)\right\}^{-1} \tag{12}$$

where

$$R_{\mathrm{a}} = \left(\Delta + \gamma\right) r_{\mathrm{a}}^{\mathrm{a}} \tag{13}$$

$$R_{\mathrm{s}} = \left(\Delta + \gamma\right) r_{\mathrm{a}}^{\mathrm{s}} + \gamma r_{\mathrm{s}}^{\mathrm{s}} \tag{14}$$

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

Soil surface resistance is expressed as:

$$r_{\mathrm{s}}^{\mathrm{s}} = \exp\left(b_1 - b_2 \frac{\theta}{\theta_{\mathrm{s}}}\right) \tag{16}$$

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 (r_a^c~~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 model (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~~ The ET is expressed as:

$$\lambda ET = \alpha_{PT} \frac{\Delta}{\Delta + \gamma}\left(R_n - G\right) \tag{17}$$

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

$$\lambda ET = \alpha \frac{\Delta}{\Delta + \gamma}\left(R_n - G\right) \tag{18}$$

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:

$$\alpha = \beta_1\left[1-\exp\left(-\beta_2\Theta\right)\right] \tag{19}$$

where ~~Θ is calculated as~~ $\Theta = \frac{\theta-\theta_r}{\theta_s-\theta_r}$.

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:

$$\lambda ET = \left(2\alpha_{PT}-1\right)\frac{\Delta}{\Delta+\gamma}\left(R_n-G\right)-\frac{\gamma}{\Delta+\gamma}\frac{\rho\left(q^*-q\right)}{r_a} \tag{20}$$

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 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~~as for the PT-FC model, we ~~modified~~ changed 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

The Bayesian model evidence (BME) of a model, *M*, is defined as (Schöniger et al., 2014):

$$\mathrm{BME} = p\left(\mathbf{D}\middle|M\right) = \int p\left(\mathbf{D}\middle|\boldsymbol{\theta},M\right)p\left(\boldsymbol{\theta}\middle|M\right)d\boldsymbol{\theta} \tag{21}$$

where **D** is observed or estimated data, $\boldsymbol{\theta}$ is the vector of parameters associated with model *M*, $p\left(\boldsymbol{\theta}\middle|M\right)$ is the prior density of $\boldsymbol{\theta}$ under model *M*, $p\left(\mathbf{D}\middle|\boldsymbol{\theta},M\right)$ is the joint likelihood of model *M* and its parameters $\boldsymbol{\theta}$. Estimating BME using power posterior estimators such as thermodynamic integration (TI) (Lartillot and Philippe, 2006) depends mainly on the calculation of the marginal likelihood $p(\mathbf{D}|M)$. 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

$$q_\beta(\boldsymbol{\theta}) = p(\mathbf{D}|\boldsymbol{\theta},M)^\beta\, p(\boldsymbol{\theta}|M) \tag{22}$$

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 (22) uses the normalizing constant Z_β to yield the normalized power posterior density:

$$p_\beta(\boldsymbol{\theta}) = \frac{q_\beta(\boldsymbol{\theta})}{Z_\beta} \tag{23}$$

such that

$$Z_\beta = \int q_\beta(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{24}$$

The above integral takes a simplified form by the potential:

$$U(\boldsymbol{\theta}) = \frac{\partial \ln q_\beta(\boldsymbol{\theta})}{\partial \beta} \tag{25}$$

thus, the integral can be directly estimated by the following way:

$$p(\mathbf{D}|M) = \frac{Z_1}{Z_0} = \exp\left\{\int_0^1 E_\theta\left[\ln p(\mathbf{D}|\boldsymbol{\theta}, M)\right] d\beta\right\} \tag{26}$$

The one-dimensional integral with respect to β is evaluated by using numerical methods by discretizing β into a set of β_k. Since there is no theoretical method for selecting β_k 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

$$\beta_k = (k/K)^{1/\varepsilon} \tag{27}$$

for k =0, 1, 2…, K. Using $\varepsilon = 0.3$ and $K = 20$ is a reasonable initial choice. By using the trapezoidal rule of numerical inregration, equation (26) is evaluated via

$$p(\mathbf{D}\,|\,M) = \exp\left(\int_0^1 y_\beta d\beta\right) = \exp\left(\sum_{k=0}^{K} r_{TI,k}\right) \tag{28}$$

such that

$$r_{TI,k} = (\beta_k - \beta_{k-1})\left[\frac{y_k - y_{k-1}}{2}\right] \tag{29}$$

and

$$y_k = E_\beta[\log p(\mathbf{D}\,|\,\boldsymbol{\theta}_k, M)] = \frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{D}|\boldsymbol{\theta}_{k,i}, M) \tag{30}$$

where n is the number of random samples of $\boldsymbol{\theta}_k$ corresponding to β_k, and $\boldsymbol{\theta}_{k,i}$ is the i-th sample.

The random samples, $\boldsymbol{\theta}_{k,i}$, are drawn by using the MCMC method implemented in the DREAM code. See Appendix B for further details on Bayesian inference and the DREAM algorithm. In the DREAM-based calculation, the Metropolis acceptance ratio is $\alpha_k = \min\left(1,[\alpha_{k,power-posterior}\alpha_{k,prior}]\right)$ 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 $\boldsymbol{\theta}_{k,old}$. 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 algorithm to sample any power posterior distribution, 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.

2.5 Traditional statistical metrics of evaluating model performance

The traditional error metrics for evaluating model performance include ~~correlation-based measures of~~ $R^2$and slope (correlation-based measures), ~~relative error measures of~~index of agreement (IA) and model efficiency (EF) (relative error measures), and ~~absolute~~ the root mean square error ~~measures of~~ (RMSE) and mean bias error (MBE) (Poblete-Echeverria and Ortega-Farias, 2009). The~~ir~~ definitions~~s~~ of the listed metrics are ~~as follows~~:

$$\mathrm{IA} = 1 - \frac{\sum_{t=1}^{n}\left[O(t) - M(t)\right]^2}{\sum_{t=1}^{n}\left[\left|O(t) - \overline{O(t)}\right| + \left|O(t) - \overline{M(t)}\right|\right]^2} \tag{$\cancel{33}3\cancel{0}1$}$$

$$\mathrm{EF} = 1 - \frac{\sum_{t=1}^{n}\left[O(t) - M(t)\right]^2}{\sum_{t=1}^{n}\left[O(t) - \overline{O(t)}\right]^2} \tag{$\cancel{34}3\cancel{1}2$}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left[O(t) - M(t)\right]^2} \tag{$\cancel{35}3\cancel{2}3$}$$

$$\mathrm{MBE} = \frac{1}{n}\sum_{t=1}^{n}\left[O(t) - M(t)\right] \tag{$\cancel{36}3\cancel{3}4$}$$

where $O(t)$ is the observation~~s~~ and $\overline{O(t)}$ is the mean observation at time t~~,~~; $M(t)$ is the modeled value and $\overline{M(t)}$ is the mean 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

The ~~re were~~PM model has five parameters g_{max}, D_{50}, Q_{50}, K_q and K_a;~~$_A$ in the PM model, and two~~ the SW model has seven parameters – the five used in the PM model and ~~additional~~ parameters~~,~~ b_1 and b_2~~,~~. ~~in the SW model.~~ The PT-FC and AA models each include two parameters, denoted ~~as~~ by β_1 and β_2 (Table 1). The prior probability density of each parameter ~~was~~ is specified as a uniform distribution with the ranges listed in Table 1. A total of 50,000 realizaitions were generated with the DREAM algorithm, which was used to estimate the posterior probability density function of each parameter with the calibration period data from DOY 154 to DOY 202. In the calculations, the chain number, N, was equal to the number of parameter~~ number~~s in the associated model. Therefore,~~ i.e.,~~ N ~~=~~is equal to 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 for calibration.~~ to set up posterior density functions for each chain.~~ In total, 40,000 ×N realizations were used to set up posterior density functions for each model. To ~~understand~~ illustrate 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~~using a different color~~ (PT-FC and AA models not shown).~~. The algorithm required about 8,000 generations to make the G-R statistic close to 1.0 ~~1.2~~ for the ~~both~~two 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. Figures 2*a*-2*e*, 2*f*-2*l*, 2*m*-2*n*, and 2*o*-2*p* show histograms of the PM, SW, PT-FC and AA models, respectively. Parameter g_{max} (Fig. 2*a*) in the PM model, parameters g_{max}, K_a, b_2 (Fig. 2*f*, 2*j*, 2*l*) in the SW model, and parameter β_1 (Fig. 2*m*) in the PT-FC model and AA model (Fig. 2*o*) were well constrained and occupied a relatively small range.– ~~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 parameters displayed a unimodal distribution and appeared approximately Gaussian. In contrast, the distributions of the other parameters differed significantly from a Gaussian distribution, as shown by the corresponding histograms. The distributions of all but one of these parameters concentrated most of the probability mass at their upper limits.~~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~~ The exception was parameter b_1 for the SW model (Fig. 2*k*), which clearly does not follow a normal distribution with most of the mass ~~obviously deviated from normality and tended to~~ concentrated 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. 2~~J~~g) but was also the upper edge-hitting parameter in the SW model. ~~In addition~~Moreover, the corresponding distributions of the same parameter in different models were slightly different~~;~~. ~~for~~ 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~~ 2a and ~~2I~~2f~~;~~, 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. It is interesting to observe that the distribution of K_a in PM model (Figure 2*e*) has a truncated distribution with highest probability mass at the upper bound, whereas the distribution of K_a in the SW model (Figure 2*j*) tends to become approximately normal. Overall, the marginal posterior probability density function of most of the individual parameters occupied only a relatively small region compared with the uniform prior distributions, and exhibited relatively large uncertainty reduction.

3.2 Performance of the models

The performance~~s~~ of each of the four ~~evaporation~~ ET models ~~were~~ was evaluated ~~during~~ over the course of 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. Table 3 summarizes the ~~S~~statistical results for the performance of the models ~~were summarized in tables as the~~using 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 in the predictions made by~~among~~ the 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 model ~~overestimates~~ overestimated 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 that of the ~~SW, PT-FC and AA~~other three models, with a slope almost equal to 1 and ~~with~~ relatively lower MBE. The SW model was ranked second, while performance of the AA model was slightly higher comparable to that of the PT-FC model~~, but slightly higher, and was ranked third~~. However, if R^2, RMSE, IA, and EF were used to evaluate ~~the model~~ performance~~s~~, 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.~~and t~~T~~he 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 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~~error metrics, 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~~arise in MBE for the four ET models. From the variations ~~of~~ in 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 indicate~~d~~ 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~~ is currently no theoretical method ~~so far~~ for selecting power posterior 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. Although adding more β_k values might improve the BME estimation, this was not done because of the computational cost. For each β_k value, at least 150,000 DREAM simulations were large enough to ensure convergence. Figure 5 ~~showed~~ shows 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 -6290.71and 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.The BME-based model ranking (from the best to the worst) is SW, PM, PT-FC, and AA. ~~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, T~~the PT-FC and AA models, which consisting ~~of~~ the same number of parameters, had similar potential patterns of evolution with ~~the~~ respect to 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.~~ The results illustrate that with the addition of parameters, the model complexity and the model performance are both increased.

4. Discussion

4.1 Parameter uncertainty analysis

With regard to the efficiency of the DREAM algorithm, the acceptance rates of the PM (15.3~~3~~%) and SW (18.9~~4~~%) models were much higher than those obtained by some Markov Chian Mote Carlo (MCMC) algorithms ~~which~~ that have been used in ~~the~~ previous studies. ~~, like 0.01902% in the population Monte Carlo sampling algorithm~~ (Sadegh et al., 2014). ~~This was is a large improvement in search efficiency, which in large part resulted resultes from its ability to sample one or groups of variable in turn. Furthermore, this the DREAM algorithmmethod ranruns multiple chains in parallel and adaptively updateds the scale and orientation of the proposal proposed distribution (Vrugt et al., 2008). Therefore, the DREAM scheme substantially improved improves not only the convergence, but also its sampling efficiency for ET models.~~ The posterior parameter bounds exhibit a larger reduction using the DREAM algorithm compared with other studies using the Metropolis–Hasting algorithm. This demonstrates that DREAM could efficiently handle problems involving high-dimensionality, multimodality, nonlinearity.

The results showed that the assumed prior uncertainty ranges from most parameters in the four models were significantly reduced. This indicates that the observed ET data contained sufficient information to estimate these parameters. Surface conductance g_s and modeled ET in the PM model are relatively insensitive to Q_{50}, D_{50} and K_q. Hence, these parameters could not be well constrained, and further relaxing the ranges for these parameters could not result in physically realistic behavior of the model. The calculation of g_s^c –in the SW model is the same as in the PM model, and thus, g_s^c and modeled ET in the SW model are also insensitive to parameters of Q_{50}, D_{50}, K_q. Therefore, these three parameters were also not well constrained in the SW model. In addition, the uncertainties present in the edge-hitting parameters, may be the outcome of model biases or EC-measured ET data errors, or the characteristic time scale of parameters governing the processes affecting ET is 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 the humidity deficit at which stomatal conductance is half its maximum value, respectively, and these parameters may change over a shorter or longer time scale than half-hours.

The ecophysiological parameter g_{max} is a variable in the g_s^c–equation in both the PM and SW models, but this parameter is sensitive to g_s^c and has a significant impact on the evaluated ET. ~~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~~ Its effects ~~were~~ is relatively independent compared to the other meteorological parameters in the models, and therefore this parameter was well specified in the PM and SW 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.~~ Parameter β_1 was well constrained in the PT-FC and AA models because it was relatively independent and did not directly relate to other observed variables.

~~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 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.~~

Parameter K_a implicitly appears in the surface conductance equation (Eq.2) in PM model and K_a is insensitive to g_s and modeled ET (Leuning et al., 2008). In contrast, K_a is contained in the equation of net radiation flux into the substrate (Eq.10) in the SW model. This parameter can explicitly partition the total available energy into that absorbed by the canopy and by the soil in the SW model. An analysis of equation (10), found that the variation of K_a could not only account for the extinction effect but also correct the energy forcing data errors. This also meant that the estimated value of K_a using calibration data was actually not just the true extinction coefficient, but also included the energy imbalance correction in the SW model. From this analysis, we could see that K_a not only involved the distribution of energy between the canopy and the soil surface but also the energy imbalance. Therefore, parameter K_a has a great influence on the performance of the SW model. This is why K_a is poorly constrained in the PM model but well constrained in the SW model. To further illustrate the insights regarding the influence of parameter K_a on the performance of the SW model, we calibrated the SW model again and reran the model with a constant value of K_a. The results showed a significant reduction in model performance when K_a was held constant. This implied that the main reason for the SW model outperforming the PM model in our study was not only the more physically rigorous structure of the SW model but also the key parameter K_a being well constrained in the SW model.

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 arose because of~~came from~~ the parameter uncertainty. However, this method ~~did not~~cannot explicitly consider errors in the input data and model structural inadequacies. This is unrealistic for real applications, and it is 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~~ metrics (such as regression slope and MBE) and squared-residual-based measures (such as R^2, RMSE, IA, and EF). The rankings of the models obtained using the same type of metric (residual-based or squared-residual-based) are similar. ~~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~~ Slope and MBE, for example, which ~~have similar results in the~~are both residual-based measures, produce identical rankings.

However, the rankings produced by metrics of different types are not the same. ~~results from different kind of measures were incongruent; for~~ For example, the PM model ~~outperformed~~ outperforms the SW model according to the residual-based ~~measures~~metrics, but the performance of the PM model ~~was~~ is worse than SW model based on the squared-residual-based measures. The comparative analysis shows a consistency between BME and the squared-residual-based ~~statistics~~metrics (hence the residual-based metrics disagreed with the BME measures).~~,~~ ~~whereas residual-based criteria were obvious disagreement with the BME measures.~~ ~~It~~ This ~~revealed~~ reveals that the more complex SW model ~~was~~ is 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~~is: 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 have 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 the 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 model with the best performance~~-behaved model~~. This ~~likely~~ probably resulted from the fact that the complex SW model ~~was~~ is indeed the most reliable model among the alternative ET models and can provide a good fit to justify its higher complexity. The SW model is a two-layer model, and simulates soil evaporation and plant transpiration separately, whereas the PM model is a single-layer model in which the plant transpiration and soil evaporation cannot be separated (Monteith, 1965). The PT-FC model ~~was~~ is a simplified ~~model~~ version of the PM model, and ~~it~~ only ~~required~~ requires meteorological and radiation information (Priestley and Taylor, 1972), whereas the AA model only ~~relied~~ relies on the feedback between actual ET and potential ET (Brutsaert and Stricker, 1979). Based on these physical mechanisms and processes that each of~~for~~ these ET models take into account, the rank order of the models ~~was~~ is 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~~ is the best performing model ~~–evaluated by~~in terms of squared-residual-based ~~measures~~metrics, which ~~resulted~~ resultes 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~~ metrics seemed to identify a reasonable rank order, this ~~had~~ has ~~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~~ is associated with these ~~measures~~ metrics and leads to relatively high values due to the squaring of the residual terms (Willmott, 1981). Furthermore, these traditional statistical ~~measures~~ metrics ignor~~ed~~es the priors, without penalizing model complexity, which ~~was~~ is 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, provide~~d~~ identical estimates of R^2 and IA. This ~~was~~ is most likely because both ~~the PT-FC and AA~~ models had the same dimension~~s~~ 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~~ would selecte~~d~~ the complex model as the best performing model. ~~Thus, we deduced that SRB measures are also problematic.~~ As for slope and MBE, the rankings produced by these~~se~~ residual-based ~~measures~~ metrics were in obvious disagreement with the one based on 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 criteria ~~s~~ provide~~d~~ an erroneous and unreliable evaluation of the model~~s evaluation~~. Therefore, the squared-residual-based and residual-based measures were not certain to provide reasonable results in terms of model ranking.

BME is a consistent model selection which tries to identify which of the models produced the observed data. Conversely, nonconsistent model selection uses the available data to estimate which of the models might be best in predicting the future data. In fact, the error metrics are essentially nonparsimonious model selection, which is a special case of nonconsistent model selection, where only the goodness of fit is used for rating models without penalizing the model complexity and thus lacking consistency for the selected model (Höge et al., 2018). The consistency between BME and the squared-residual-based metrics only indicates that the optimal model evaluated by BME would also provide the best predictions, and thus consistent model selection should also be asymptotically efficient (Leeb & Pötscher, 2009; Shao, 1997).

4.3 Analysis of model-data mismatch

Conceptual and structural inadequacies of the hydrological model ~~and~~together with measurement errors of the model input (forcing) and output (calibration) data introduce~~d~~ errors in the estimated parameters and model simulations (Laloy, ~~2014~~2015). Hydrological systems ~~were~~ are 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 raise~~occur~~d~~ 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~~ is described using equations that can only capture parts of the complex natural processes and ~~the~~ any ET model ~~structures were~~is 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~~ was 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 day~~time~~, whereas it was about 0.25 ~~during night-time~~at night. However, the measured ET only included vertical flux and not horizontal flux, leading to the measured ET being lower than that of ~~modeled~~ ET predicted 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~~ makes 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 estimate 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 that the PM model combined with canopy resistance overestimated maize ET during the partial and dense canopy stages by 16% and 13%, respectively. Moreover, in a study of ET in vineyards, Leuning (2008) found that the PM model coupled with canopy resistance overestimated ET during the entire growth stage by 29%.

The ~~estimated~~ estimates for ET ~~for~~ produced by the PT-FC and AA models ~~was~~ were 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, t~~The PT-FC and AA models consistently underestimated ET, especially during the late maturity stage.~~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) The l~~L~~ack of mechanistic representation of rainfall interception in ET models probably ~~also~~ led to inaccurate simulation ~~on~~ for periods soon~~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 includes 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 is independent of relative humidity. However, the similar statistical results and similar degrees of MBE scatter indicate 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 show that the model-data disagreement is not caused 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 think that the consistent underestimation of ET by the PT-FC and AA models results primarily from conceptual and structural inadequacies, energy imbalance, and soil water stress. Although the PM and SW models share a common theoretical basis and the PT-FC model is a simplification of the PM model, these models perform 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 estimates of ET by the PM and SW models are characterized by symmetry and a low degree of scatter, but the PT-FC and AA models exhibit consistently 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, some studies reported that the PT-JPL (Zhang et al., 2017) and the AA model showed an overall poor performance (Zhang et al., 2017). While 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 but also for different meteorological, physiological and soil conditions. Moreover, the performance is also related to the stage of crop growth. Note that these conclusions about the ET models evaluation are derived from traditional error metrics rather than those based on BME model selection. It would be desirable to use available data from other study areas or from other crops for BME-based model selection to confirm whether the SW model is the optimal model under other conditions.

Overall, combined with the parameter uncertainty analysis described in Section 4.1, we conclude that energy imbalance and energy interaction between canopy and soil surface have a greater impact on the model performance. And thus, explicitly treating of energy error, and incorporating the elements of existing hydrologic theory about energy interaction between canopy and surface or conceptually correcting the energy interaction are a practicable option for model improvement and application.

5. Conclusions

This study illustrated the application of the Bayesian approach ~~for~~ on the statistical analysis and model selection of four widely used ET models. The results showed that the DREAM algorithm successfully reduced the assumed prior uncertainties for most of the parameters in the four models. In the model calibration, the key parameters which had a significant influence on ET simulations were well constrained. The main reasons for the outperforming of SW model were its physically rigorous structure and the extinction coefficient parameter, which is sensitive and has a significant impact on the performance of the model, being well constrained. –

~~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.~~ BME is a consistent model selection to~~which tries to~~ identify thes best fitting to the observed data.Although the squared-residual-based metrics, including R^2, IA, RMSE, and EF, produced a ranking identical to that of BME, it must be noted that these squared-residual-based metrics do not allow using prior information and do not penalize the model complexity when comparing the models. Therefore, some cautions are needed when using these statistical methods to compare different models.

The model–data ~~mismatches~~ discrepancies were analyzed to facilitate model improvement after ~~using~~ Bayesian model calibration and comparison. The results indicate~~d~~ that ~~model–data mismatches~~the discrepancies arose ~~are~~ mainly as a result of energy imbalance caused by anisotropic turbulence, additional energy induced by advection processes, the absence of a mechanistic representation of the physiological response to plant hydrodynamics and the energy interaction between canopy and surface. Among these causes, energy imbalance and additional energy are related to forcing data errors rather than to an unreasonable model structure. Thus, understanding the process of the physiological response to plant hydrodynamics and the interaction between canopy and surface is essential for improving the performance of evapotranspiration models. Overall, the applications of Bayesian calibration, Bayesian model evaluation and analysis of model–data discrepancies in our study, provide a promising framework for reducing uncertainty and improving the performance of ET models. It would be desirable to confirm whether the SW is the optimal model using data of other crops.or other climate regions.

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 and Nos.41702244. 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

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, 2000.

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, 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, 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, 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, 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, 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, doi:10.1002/2015WR017871, 2016.

Höge, M., Wöhling, T., and Nowak, W.: A primer for model selection: The decisive role of model complexity. Water Resour. Res., 54, 1688–1715, doi.org/10.1002/2017WR021902, 2018.

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, doi:10.1002/2014WR016395, 2015.

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

Leeb, H., and Pötscher, B. M.: Model selection, Berlin, Germany: Springer., pp, 889–925, doi.org/10.1007/978-3-540-71297-839, 2009.

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, 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, 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.: Comparison of several surface resistance models for estimating crop evapotranspiration over the entire growing season in arid regions. Agric. For. Meteorol. 208, 1-15, 2015.

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, 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 thermodynamic integration method and comparison with several other numerical methods. Water Resour. Res., 52(2), 734-758, 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, 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, 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., 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, 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, 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, doi:10.1002/2014WR016062, 2014.

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

Shao, J.: An asymptotic theory for linear model selection, Statistica Sinica, 7(2), 221–242, 1997.

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, 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.

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.

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.

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. Journal of Geophysical Research: Atmospheres, 122(1), 230-245, 2017.

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.: 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.

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)
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 ($m^3\ m^{-3}$)
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:

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

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

$$p(D \mid M) = \int p(D \mid \boldsymbol{\theta}, M)\, p(\boldsymbol{\theta} \mid M)\, d\theta \quad \text{(A2)}$$

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

The likelihood function, $p(D|\boldsymbol{\theta}, M)$, 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

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

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

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

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

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

		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); Barton. (1979)
β_2	empirical constant					0.1	10	

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

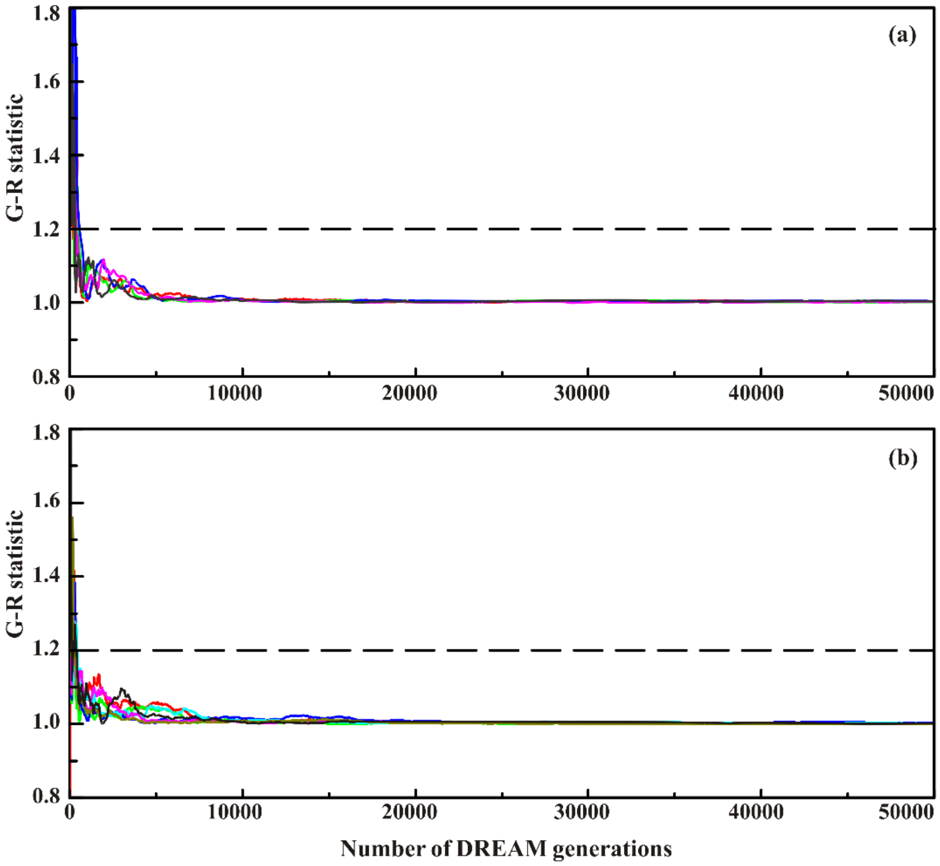

Figure 1 Trace plots of the G-R statistic of Gelman and Rubin (Gelman and Rubin, 1992) using DREAM for the PM model (a) and (b) the SW model. Different parameters are coded with different colors. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution.

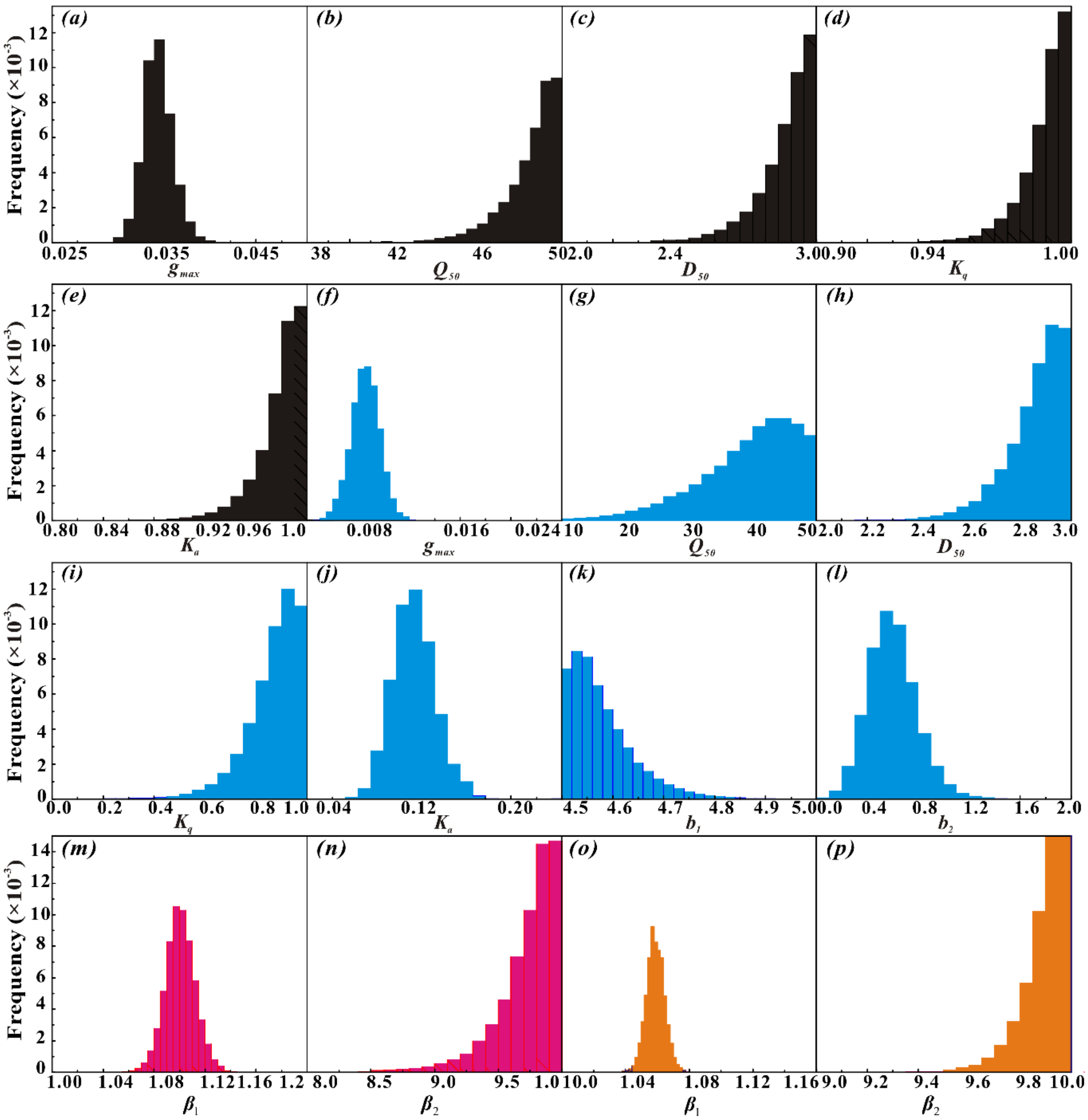

Figure 2 (*a*)-(*e*), (*f*)-(*l*), (*m*)-(*n*), and (*o*)-(*p*) show histograms for the PM (black), SW (cyan), PT-FC (magenta) and AA (orange) models, respectively. These histograms are constructed from all chains for each model and a total of 40,000×N realizations are simulated using DREAM. The x axes represent the prespecified limits of the parameters.

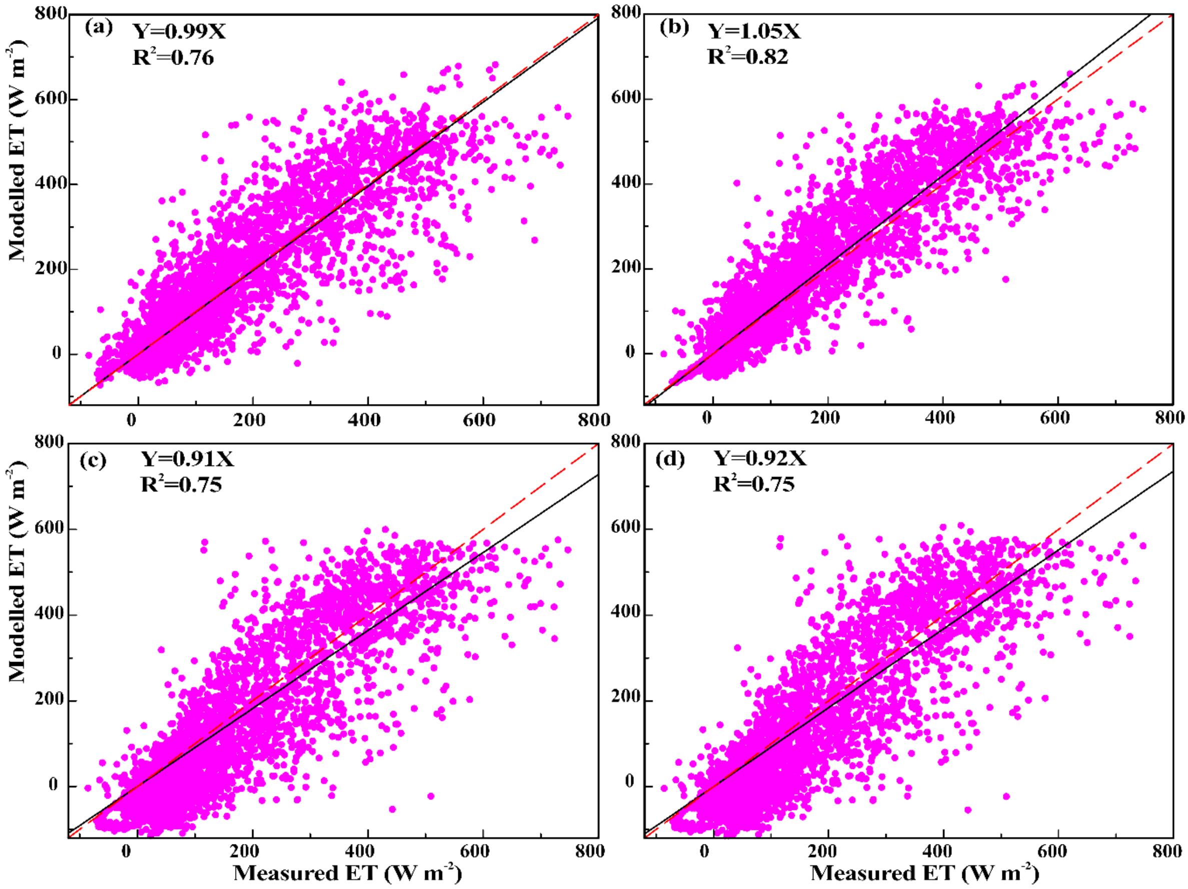

Figure 3 Regressions between measured and modeled half-hourly ET values produced by different models from DOY 154 to DOY 270: (a) PM, (b) SW, (c) PT-FC and (d) AA. The regressions are: Y = 0.99X (R^2 = 0.76), Y = 1.05X (R^2 = 0. 82), Y = 0.91X (R^2 = 0.75), and Y = 0.92X (R^2 = 0.75) for the PM, SW, PT-FC and AA models, respectively.

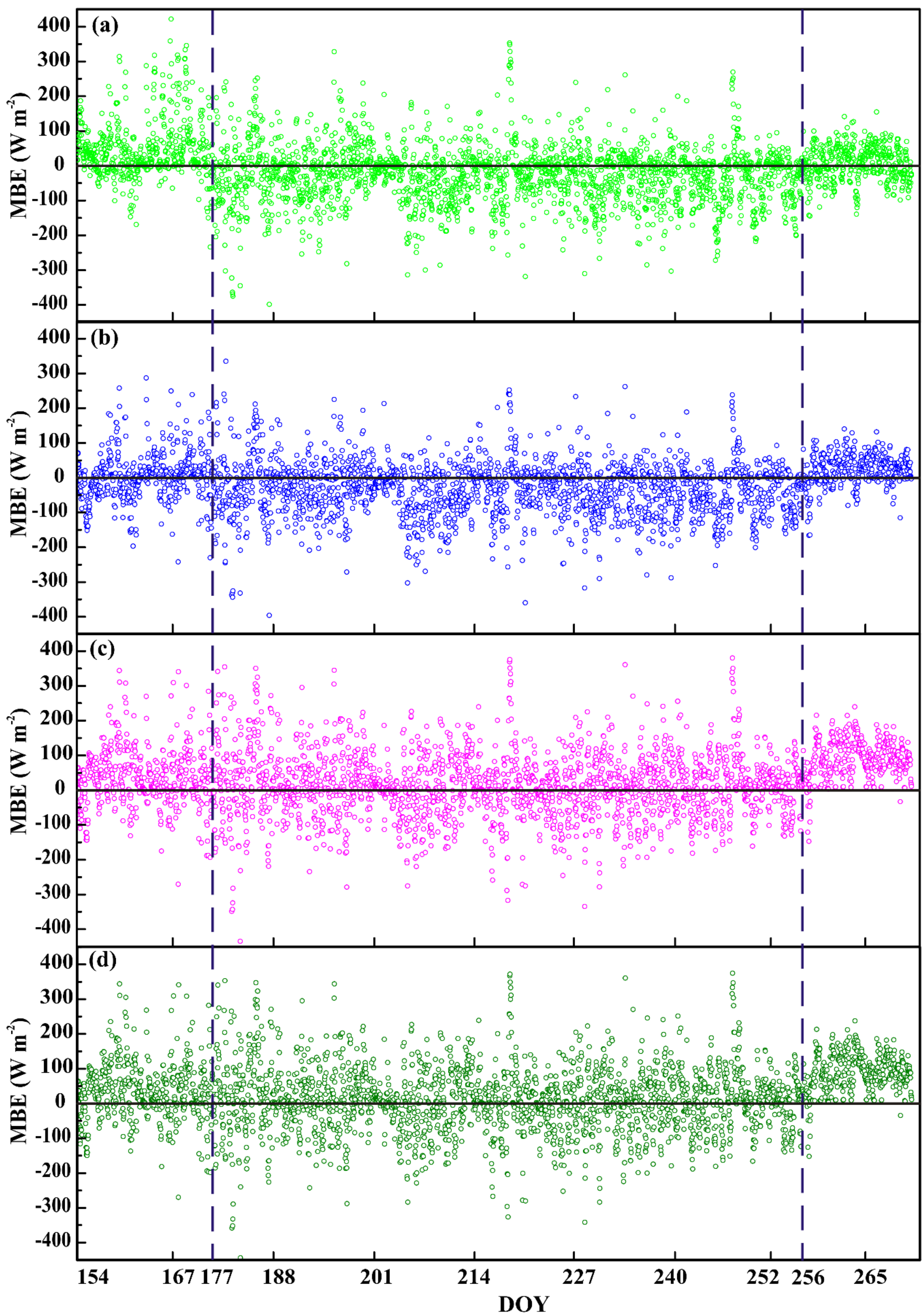

Figure 4 Mean bias error (MBE) of predicted and observed ET values for (a) PM, (b) SW, (c) PT-FC and (d) AA models from DOY 154 to DOY 270. Parameters used for prediction are estimated by DREAM with the dataset for the calibration period from DOY 154 to DOY 202.

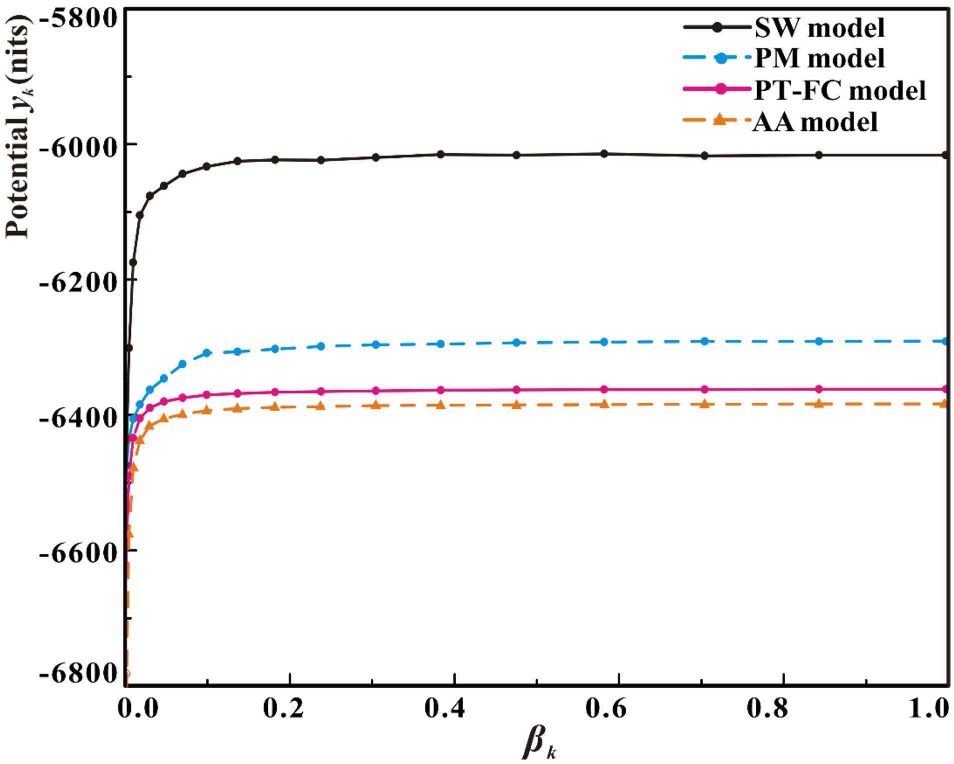

Figure 5 Variation of the mean posterior expectation of the potential y_k with β_k for the PM, SW, PT-FC and AA models.