Referee #1:

General comments:

1. Response to 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 improve 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. Response to 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: Yes, this comment is quite useful. We could change the original objective (1) and (2) to "(1) to calibrate ET model parameters using the diffeRential evolution adaptive metropolis (DREAM) algorithm (Vrugt et al., 2009); (2) to identify which parameters had a greater impact on the model performance and to explain why the selected optimal model performed best".

3. Response to 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 will add more discussions in 4.3 "Analysis of model-data mismatch".

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

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 will be edited by the professional editing services.

Specific comments:

1. Response to 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. The ET model parameters were calibrated using the entitled differential evolution adaptive Metropolis (DREAM) algorithm; the optimal model was selected using Bayesian model evidence (BME), which was implemented using the mathematically rigorous thermodynamic integration method. The estimated parameters were analyzed to identify which parameters had a greater impact on the models' performance and explain why the optimal model performed best. The discrepancies between observations and model estimates were evaluated using traditional error metrics in order to analyze the shortcomings of the different ET models and find ways to improve their performance. The results indicate that DREAM can effectively infer the model parameters. Although the SW model with seven parameters is more complex than the other three models, the BME criterion still selects SW as the best model. This is because the structure of the SW model is more physically rigorous, its parameters have greater impact and their sensitivity is well constrained. Moreover, the results show that energy imbalance and energy interaction between canopy and surface can bias the model estimates significantly.

2. Response to 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 changed this statement to "We anticipate that our study could provide a basis for improving ET models".

3. Response to 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 originally considered there are several competed methods can select the best-performing model. But as the reviewer said, this is "well-known", thus we decided to delete this sentence.

4. Response to 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. We are sorry that our expression was not very clear. We will modify the words and make it clearer.

5. Response to 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: Yes, this referred the 40,000 samples from one chain. Total have 40,000 multiplied by *N* (chain number) samples from all chains.

6. Response to 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 comments. We decided to change 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 will delete the 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".

7. Response to 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 will redraw the Figure 1. This can see in the modified Figure 1 .

8. Response to 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: Thanks for the comment. We have redraw the Figure 2 as in the attached file.

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.

9. Response to 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 comment. We will revise this sentence.

10. Response to 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 will revise this sentence and delete all the confusing words.

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

Response: We will correct these type errors.

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