Response to Reviewer 2

[Cover Letter]

Dear Reviewer,

We appreciate you for spending time to review our paper and providing some valuable comments. It is your valuable and insightful comments that led to possible improvements in the current version. The authors have carefully considered the comments and tried our best efforts to address every one of them. However, some revisions may still cannot meet your high standards. The authors welcome further constructive comments if any. We provided the point-by-point response first and will provide the updated version of the paper after proofreading complete.

Sincerely,

Bo Huang, PhD

bohuang@cuhk.edu.hk

Professor, Department of Geography and Resource Management

The Chinese University of Hong Kong

[General Comment] The paper addresses the importance of model evaluation and presents a robust method for evaluating the results from urban climate simulations. Overall, the paper is clear and well structured. The discussion on natural gap, observation bias and model bias is substantial, highlighting the problems existing in current modeling practices that the climatological modelers should pay more attention to. The study is valuable to be published in a high impact journal. I would suggest a minor revision in which the authors should focus more on evaluation framework and clarify some technical points.

Response: The reviewer made constructive comments to improve the presentation and structure of the paper. We better organized the presentation of the proposed evaluation framework by including a clear workflow of the evaluation framework, more justification of PSS theory and other tools, and a summary table of evaluation results in our case study. With respect to the definition of 'acceptable', we discussed it and summarized a framework of the practical grading guidelines, which shall be further refined given the proposed evaluation tools being applied in many other case studies. Moreover, thanks for your interest in the developed high-resolution urban surface data. We are preparing another paper on it in which we compared the modeling

results using the coarse urban land surface data provided by the WRF ARW model and the newly developed high-resolution urban land surface data. The paper should come out soon. Furthermore, regarding the selection of schemes for the physics components, provided more details in the later version.

Major Comments:

[Comment 1] The focus of this paper should be the model evaluation. The authors may strengthen the introduction and discussion of the evaluation framework in the following aspects:

(1) Presentation of the evaluation framework: the authors should summarize and present the evaluation framework in a visualized and more straightforward way (for example, using a workflow diagram).

Response: Thank you for your suggestions. We added a detailed explanaion of the proposed model evaluation framework in Section 2.2. We also added a workflow diagram for better presentation of the proposed model evaluation framework. We hope the extended explanations made the proposed framwork easier to understand.

(2) Justification for the evaluation tools: the authors should introduce more PSS theory and explain why it is suitable to evaluate the model for urban climate simulations. The same as the PDF analysis and other evaluation tools.

Response: The importance of examining climate statistics other than climate means is not new (Katz and Brown 1992; Boer and Lambert 2001). The descriptive statistics are useful in providing aggregated information on the distribution of the attributes, but they can be very misleading since very different distributions can lead to similar descriptive statistics, and these aggregated metrics can be sensitive to outliers. Therefore, we examine not only the descriptive statistics but also metrics regarding the statistical distributions of modeled and observed meteorological attributes. The advantages of PDF and PSS for climate statistics have been discussed by Perkins et al. (2007). We have revised accordingly in Section 2.2.

(3) Interpretation of the evaluation results: the authors kept using “acceptable” to describe the results. But how to define “acceptable”? What is the value of PSS would be considered as “not acceptable”? To make it a complete framework, the authors should provide guidelines to evaluate the results from the model evaluation.

Response: Thanks for your great suggestions. A reasonable definition of "acceptable" would improve the novelty of this manuscript.

We summerized the 72 monthly analysis for 6 meteorological attributes in our case study. The PSS values generally followed a normal distribution ranging from 0.444 to 0.886 with an average of 0.660 and a standard deviation of 0.098. Therefore, the PSS values are larger than 0.500 with a probability of 95%.

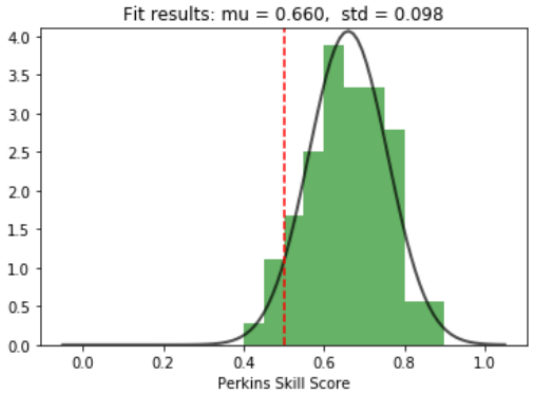

Figure 1. Histogram of the Perkins skill score for 72 monthly PDF analysis. A normal distribution was fit. The red dashed line indicate the lower bound for 95% confidence (PSS=0.500).

Based on our results, we define 2 criteria for "acceptable" high-resolution urban climate simulations: 1) Yearly average PSS $\geq$ 0.550; 2) For each meteorological attribute, PSS $\geq$ 0.500 with a confidence interval of 0.500.

Compared to the case studies in Perkins et al. (2007), the lower bounds for PSS in our standard was lower, which is due to the increased resolution in our simulations. We are fully aware that, despite the sophisticated analysis we have conducted on different spatial and temporal scales, it is difficult to define 'acceptable' using results from one case study. In this vein, the proposed standard of 'acceptable' for high-resolution urban climate simulations based on our case study was meant to be the starting point and to be improved by future case studies using the proposed model evaluation framework.

(4) Intervals in PDF analysis: the authors use intervals of [-1, 1], [-2, 2], [-3, 3] for all variables in the PDF analysis. However, the significance of 3 degree in temperature change should have higher impact than 3 millimeter in precipitation. The authors should consider how to choose reasonable intervals for different variables.

Response: Thank you very much for your comment. Indeed, the same intervals have very different meanings for different meteorological attributes. Therefore, we are planning to use the standard deviations as intervals of the PDF analysis instead of fixed intervals.

(5) Selection of variables: the authors should state the rationale for choosing variables for model evaluation in your case study.

Response: We included all meteorological variables that are meaningful for urban climate analysis in the model evalutaion.

(6) Next steps: the authors should discuss the drawbacks of the proposed evaluation framework and provide suggestions for future research. It would be a plus if the authors provide the source codes and original datasets using in the model evaluation.

Response: Thank you for your comment. We added more discussions on drawback of this study and possible future directions in the Conclusions. We are planning to open the entire dataset in another paper specifically on the development of the high-resolution urban land surface data under review.

"This study is not perfect. The effects of the selected physical components on the evaluated modelling accuracy is not clear, which requires further control experiments. Also, the effects of the refined urban land surface datasets on the evaluated modelling accuracy also requires further discussions."

[Comment 2] Although the inputs and setups in the modeling are critical to the model results, however, they are not the emphasis for this paper, and thus the modeling details should be listed in the appendix. On the other hand, a table of summarizing the evaluation results should be presented.

Response: Thank you for your advise. We are including Section 3 in Supplementary Material.

[Comment 3] Here are some suggestions the authors may take into consideration for their future research by applying their proposed evaluation method in investigating the model components and setups.

Response: Thank you very much for your suggestions. We definitely agree with you that many more can be done and we will continue to thrive in this direction.

(1) New developed urban data: the authors developed four new sets of high-resolution urban data for modeling urban climate. What impact they have on the model results? Do they improve the overall performance of the model? If so, how much the improvement?

Response: Thank you for your comment. That is a great point for future research. Acturally, we have another manuscript under review focused on the development of the high-resolution urban land surface dataset and its effects on the reliability of climate simulation. Please refer to Archive.

We can briefly introduce results from the other paper. People would naturally expect more accuracte modeling results with more accuracte input data. Unexpectedly, we find that high-resolution urban land surface datasets could either increase or decrease the evaluated reliability of simulation results, which is probably why not all modelers refine the land surface input data before the simulation. We believe the reason for this phenomenon is due to inperfect model and inperfect model evaluation methods. First, imperfectness in the detailed physical processes included in the model. This is the root why more accurate input data does not necessarily lead to more accurate simulation results. Second, it's inperfect to compare the grid-based simulation results with point-based observations. Moreover, the evaluated reliability cannot be compared across scales since high-resolution simulation contain many more details and will naturally decrease the evaluated reliability.

Nevertheless, even the decrease in model evaluation metrics does not mean that the simulation results are less accuracte. We argue that providing more accurate input data into the model is the only way to prevent the 'garbage in, garbage out' effect, motivate us to refine the model itself and model evaluation methods, and lead us towards better modeling practice.

(2) Schemes of physics components: How to choose the schemes for each component? Would the selection of schemes have impacts on PSS scores?

Response: That is another great point for future research. The interactions among selected physical components are very complex and so it can be difficult to provide solid foundations for the selections. One way out is to compare the accuracy of simulation results using different combinations of physical components. However, there are three possible challenges: 1) many variables are involved in this process, including physical components, model parameters, and spatial-temporal resolutions, which makes a very large solution space. It would be compotationally expensive to try every combination; 2) Evidence from a certain study area and time period may not be transferable to other study areas or time periods; 3) proper model evaluation metric. Comparing grid-based modeling results with point-based simulations are

naturally biased. Better evaluaed accuracy does not necessarily mean better quality of simulation results.

Therefore, we think some evidence can definitely be provided from experiments using specific study area, time period, model parameter, and spatial-temporal resolution. But it would be difficult to provide more general insights for the seleciton of physical components. Theoretical discussions and the ‘try and error’ process are still vital in refining such selections.