Response to Reviewer 1

[Cover Letter]

Dear Reviewer,

We appreciate your precious time in reviewing our paper and 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 to address every one of them. We hope the revisions meet your high standards. The authors welcome further constructive comments if any. We provided the point-by-point response as below. Modifications in the manuscript are highlighted in red.

Sincerely,

Bo Huang, PhD

bohuang@cuhk.edu.hk

Professor, Department of Geography and Resource Management

The Chinese University of Hong Kong

[General Comment] This study evaluates performance of the WRF model in terms of high-resolution urban climate modelling over an area encompassing two big cities, Shenzhen and Hong Kong. The chosen area of Shenzhen is heavily urbanized but only a small part of Hong Kong is urbanized. Perkins skill score is used as a major evaluation method throughout the evaluation. The authors argue that their study has proposed a methodological framework for evaluating model performance in high resolution urban climate simulation. I think this work is useful and has provided some information about high-resolution urban climate modelling applied to south China. I very much appreciate the authors' efforts to pursue this kind of modelling work. However, I feel that the manuscript in the current form cannot be accepted for publication. At a minimum, I would suggest some necessary revisions to make the paper publishable in the journal. But to engender a stronger paper, I feel that more extensive work might have to be done. I will leave it to the editor to decide whether such extensive work is required.

Response: The article is in pertinent response to the increasing presence of ambiguous or careless modelling practices in urban-scale climatology. It intended to state the necessity of model evaluation in urban-scale climatology modelling, draw attention within the community of

urban climate modellers, and be a kick-off in reducing these window-dressing-like modelling practices. Therefore, the purpose of this paper is to remind modellers of the necessity of model evaluation in the urban climate modelling practices rather than helping to improve the model. Moreover, the modeller should conduct a systematic model evaluation to establish the trustworthiness of the new findings from an urban climate modelling since the model cannot be verified or validated. Furthermore, we reminded that the modeller should be cautious to conclude quantitative conclusions because it is impossible to differentiate the natural gap, observation bias, and model bias in the difference between observations and its corresponding modelled results. To sum up, we are confident that this paper is important to the urban climate modeller community as it points out the pain points, that is, model uncertainties affect the trustworthiness of the new findings and it is impossible to identify the uncertainties of model completely.

[Major Comment 1] The introduction should be reformulated with greater care. The authors should survey the literature more thoroughly. Only a few papers are mentioned in the introductory section. I suggest the authors give a good overview of the existing studies on the topic and point out the limitations of the past studies and challenges/ constrains. Identifying a gap or proposing a new method as well as outlining the contributions of the study is also helpful.

Response: We added some new related literatures in Section 1 to emphasize the importance of model evaluation in urban climate modelling and the fact that modellers paid minimal attention in their modelling practices. Moreover, we identified the systematic framework for model evaluation as the research gap in the urban climate modelling community and outlined the values of this paper in Section 1. (Pg2, Ln27-34)

[Major Comment 2] The data and methodology section should be structured in a more logical way. I think the authors could place model description and experiment/model setup before evaluation method. Overall, both section 2 and 3 are a bit confusing. The introduction of the model is lacking. The authors should clearly articulate what has been done and how it has been done. This can aid the readers in understanding the experiment setup/design.

Response: We revised Section 2 to improve clarity and provided more information about model description in Section 2.1 [Pg3, Ln2-23] and set-up in Section S4 of Supplementary Material. Moreover, we will submit another paper to describe all details about the high-resolution urban climate modelling inlcuding suggestions for modelling process, the design of the atmospheric model, model set-up, primary data processing, and a framework for quality assurance.

[Major Comment 3]

In section 2.1, more details about the new dataset developed by the authors should be offered.

Response: We provided more details about the developed land surface dataset in Section S2 of Supplementary Material. Moreover, we will submit another paper to provide all details about this urban land surface dataset later.

The reasons for focusing on the simulations in the year of 2010 should be discussed.

Response: We selected the year of 2010 since it was the latest year that a complete government-initiated land survey was conducted, which provided access to high-quality field-surveyed land cover data that is crutial for the climate simulation. It requires various data sources for the development of the new land surface dataset, high-resolution urban climate simulation and model evaluation, and we have datasets available aound the year of 2010. We mentioned it in the revision of the paper [Pg3, Ln11-12].

In section 2.2, more details should be provided as to the four-day segment simulations.

Response: We provided more details about four-day segments in Section S3 of Supplementary Material.

Did the model read in restart files every four days to continue the simulation?

Response: No. Each four-days simulation segment is a separated simulation.

How may a different simulation strategy affect the modelling results?

Response: Different simulation strategies is associated with the different spin-up method, which affect the modelling results. We added a small discussion about it in Section S3 of the Supplementary Material.

In section 2.3, instead of just giving two tables, I think more detailed descriptions of the data should be given. How are the comparisons between model output (grid points) and observations (stations) made?

Response: We already discussed these scomparisons in Section 4. Moreover, we added some details about the comparison in Subsection 2.3 [Pg4, Ln14-19]. Furthermore, we would like to provide the source codes of the evaluation software packages to the readers for easy replication.

Representativeness of the observations and potential biases should be discussed.

Response: We added more details about the observation datasets in Section S5 of the Supplementary Material.

The authors should also indicate the reasons for choosing evaluation variables.

Response: We added reasons for choosing evaluation variables in Subsection 2.3 [Pg4, Ln9-12].

[Major Comment 4] In section 2.4, no references are cited regarding the Perkins skill score. Is this a suitable method for this study? There should at least be some discussion. Authors should also discuss whether this method is suitable for all the variables evaluated in the study.

Response: We conducted a small discussion about the evaluation tools in Subsection 2.2 [Pg3, Ln24 – Pg4, Ln4].

[Major Comment 5] In section 3, choosing of the parameterization schemes needs discussion.

Response: We conducted a small discussion of the selection of parameterization schemes in Section S4 of Supplementary Material.

[Major Comment 6] I think the authors should tune down many of their arguments throughout the paper to avoid overstating (e.g., P2L25-26). For example, I don't see any strong methodological framework being discussed and described in the text.

Response: We enhanced the description of methodological framework to support our statement. We add a subsection (2.2 A Methodological Framework for Urban Climate Model Evaluation) to include more details about the methodological framework [Pg3, Ln24 – Pg4, Ln4].

[Major Comment 7] I have the impression that the authors have been too obsessed with 'good results' when evaluating the model's performance. Discussing 'good results' and 'bad results' at the same time, in my opinion, is fair. It's perhaps more important to identify areas for improvements.

Response: This manuscript intended to state the necessity of model evaluation of urban-scale climatology modelling and to provide a methodological framework of model evaluation to help modellers to establish the trustworthiness of modelling results, and accordingly it focused on the modelling performance rather than to help the model developers improving the model. We added an explanation in Section 1 to emphasize the focus of this paper [Pg2, Ln28-35].

[Major Comment 8] The structure and writing are too repetitive in section 4. This is also true for the figures. The number of figures may be reduced.

Response: We did our best to rewrite Section 4. Moreover, we moved some figures to Supplementary Material for reducing the number of figures in the paper.

While the focus of the paper as stated in the paper is on the urban climate simulation, evaluation seems to be applied to also the vast rural regions. The authors should clarify this.

Response: Yes. The methodological framework of model evaluation also can be applied in the local scale climate simulation wherever in urban or non-urban areas. We added an explanation in Section 5 [Pg11, Ln24-25].

I suggest the authors focus on the most important aspects of the urban climate simulation. I would suggest some points (see following) for the authors to consider and they should further develop a better evaluation framework.

Response: Thank you very much for your suggestions. We added a subsection (2.2 The Methodological Framework for Urban Climate Model Evaluation) to describe more details about the methodological framework, which included a theoretical explanation to the statistic tools applied in model evaluation[Pg3, Ln24 – Pg4, Ln4]. Moreover, we added Subsection 2.4 , which included a graphical presentation of the workflow of model evaluation, the guideline for checking the descriptive statistics figures and the grading guidelines for PSS and PDF of the difference [Pg5, Ln2 – Pg6, Ln6].

-Some basic ability of the model such as spatial distribution temperature/precipitation and diurnal cycles of temperature must be assessed.

Response: The difference in the surface temperature between in urban and non-urban areas (spatial distribution of temperature) had be assessed in Figure 9. The difference of precipitation between in urban and non-urban areas is not significant. The diurnal cycles of 2-meters air temperature had be assessed in Figure 3.

- The weather and climate variability in the study area is strongly associated with the monsoon flow. So the investigation of the simulation of precipitation and temperature is rather important. Both the spatial distribution (not found in any of the figures in the paper) and temporal variability should be considered.

Response: We agreed that the climate variability in the study area is strongly associated with the monsoon flow. However, the monsoon flow is a mesoscale meteorological behaviour and so it is not associated with the spatial distribution of precipitation and temperature at the local scale. The spatial distribution of temperature is strongly associated with the local land surface attributes. Therefore, we added some discussions in Subsection 5.3 about the relationship in the spatial distribution between 2-m air temperature and land surface temperature [Pg11, Ln5 – 25]. Moreover, we agree that seasonal variations in temperature and precipitation are associated with monsoon flow, especially precipitation. Therefore, we added some discussions in Subsection 5.3 on the relationship between the monsoon flow and the seasonal variation of precipitation, and the relationship between the monsoon flow and the seasonal variation of 2-m air temperature [Pg11, Ln5 – 25].

In particular, the authors may identify some strong urbanization impacts on the precipitation (e.g., precipitation maxima) and temperature (e.g., urban heat island). The model's ability to capture these effects is essential.

Response: Observational data before and after the urbanization process are needed to evaluate the urbanization impacts on the precipitation and temperature. We cannot provide these evaluations because we don't have these observation data. However, we added some discussions in Section 5.3 on the relationship between the spatial distribution of the 2-m air temperature and the land surface temperature, and also the relationship between the spatial distribution of precipitation and land surface temperature [Pg11, Ln5 – 25].

In addition, simulation of sea breeze, wind distribution, boundary layer variability, and stability of the atmosphere should be examined.

Response: We agree that the land-sea breeze exists in the coastal city, and sowe provided a discussion about the modelled land-sea breeze in Subsection 5.3 [Pg11, Ln5 – 25]. These modelled meteorological features (boundary layer variability and atmospheric stability) cannot be examined by the observation due to the unavailability of corresponding observation data. Examining modelled meteorological features is meaningless without comparison with observations. Therefore, we didn't provide the examination of these two meteorological features.

The impact of urbanization on the air quality may also be discussed.

Response: This study focused on providing a methodological framework for the evaluation of urban climate models. The impact of urbanization on air quality is another big topic beyond the research scope of this study.

- The evaluation can be done separately for different seasons. The evaluation should focus on the most important aspects of urban climate/weather.

Response: Actually, all figures includes the information of monthly variations in this paper, while some also show the seasonal variations. Moreover, we emphasize that the model evaluation should focus on the comparison between the modelled variables with its corresponding observed ones. Furthermore, we added a small discussion about it in Subsection 5.3 [Pg11, Ln5 – 25].

- The scientific value can be enhanced if the authors can demonstrate how the model behaves in simulating the extreme precipitation events or heat wave/cold surge events, and How and to what extent these events may be related to the urbanization.

Response: Thank you very much for your suggestions. However, our study focused on reminding the urban climate modeller of the importance of model evaluation and establishing the trustworthiness of modelling results. We also provided a methodological framework of model evaluation, and so we didn't put too much effort on the modelling performance of simulating the extreme events. In this revision, we added some discussions about the capabilities on the simulations of the extreme events on Sections 5.3 [Pg11, Ln5 – 25].

- The model's performance between different regions in the study area and between rural and urban regions can also be compared.

Response: Thank you very much for your suggestions. We added more figures on the model's performance in urban and non-urban areas in Section S6 of Supplementary Material.

[Major Comment 9] The figures can be better designed and drawn. Captions of the figures should provide more information. The language could also be improved.

Response: Thank you very much for your suggestions. We did our best to improve the language and the figure captions.

[Minor Comment] Minor comments:

The authors should check carefully the use of words and sentences throughout the paper. I suggest some serious edits/revisions. I list only some of the examples. P1L15: add 'have' before paid. P1L26-29: Please split the long sentence. P1L37: place 'into account' immediately after 'take'.

Response: Thank you very much for your suggestions. We did our best to check the paper, corrected the language errors and rewrote the long sentences to improve the readability.

Response to Reviewer 2

[Cover Letter]

Dear Reviewer,

We appreciate your precious time in reviewing our paper and 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 to address every one of them. We hope the revisions meet your high standards. The authors welcome further constructive comments if any. We provided the point-by-point response as below. Modifications in the manuscript are highlighted in red.

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] 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 [Pg3, Ln24 – Pg4, Ln5]. We added Table 1 for better presentation of the proposed model evaluation framework, in which we included three temporal persepectives, entire period, monthly, and daily, and two groups of tools, descriptive statistics and statistical distributions. Moreover, we presented the workflow for model evaluation in Section 2.4 [Pg5, Ln2 – Pg6, Ln7]. 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 also been discussed by Perkins et al. (2007). We have also updated the manuscript accordingly in Section 2.2 [Pg3, Ln24 – Pg4, Ln5].

(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. We agree that 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%.

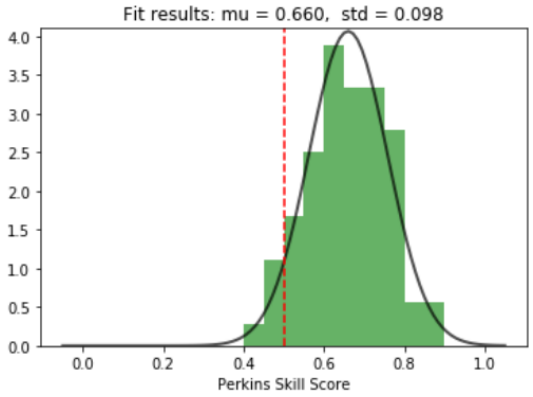

Figure 1. Histogram of the Perkins skill score for 72 monthly PDF analysis. A normal distribution was fit. The red dashed line indicate the lower bound for 95% confidence (PSS=0.500).

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 95%.

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.

We added a guideline for PSS grading in Section 2.4 [Pg5, Ln2 – Pg6, Ln7].

(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 used the standard deviations as intervals of the PDF analysis instead of fixed intervals. Moreover, we provided a guideline for specifying the interval in Section 2.4 [Pg5, Ln2 – Pg6, Ln7].

(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. We added a explanation in Section 2.3 [Pg4, Ln8 – Pg4, Ln19].

(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 Section 5.4 [Pg12, Ln15-18]. 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.

[Comment 2] 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. Regarding the modeling details (Section 3), we agree with comments from a previous reviewer that modeling details such as the input data processing and physical schemes used in our simulation are necessary to be included in the manuscript since they could have a significant impact on the simulation results. Therefore, we kept brief modeling details in Section 3 and included more in Section S4 of Supplementary Material.

[Comment 3] 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. We added some ideas for future research in Section 5.4 [Pg12, Ln15-18].

(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.