Response to Reviewer 4

[Cover letter]

Dear Reviewer,

We appreciate your devoted time in reviewing our paper and your valuable comments which enabled improvements in the current version of the manuscript. The authors have carefully considered all 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. Revisions have been made to update the manuscript (highlighted in red) and a detailed point-to-point response is provided below.

Sincerely,
Bo Huang, PhD
bohuang@cuhk.edu.hk
Professor, Department of Geography and Resource Management
The Chinese University of Hong Kong

[General Comment] The motivation and objectives of the study are of prime importance, and this revision is a better-organized and streamlined version of the original submission. However, in my opinion, the work still lacks in rigor and depth to be granted publication in GMD (see MC1).

Response: Thank you for your comments. It was good news to us that the last revision was improved in your opinion. We have gone through your new comments carefully and tried our best to address them one by one. We hope the manuscript has been improved accordingly. Model evaluation is an essential but overlooked topic. The purpose of this paper is to remind urban climate modelers of the importance of model evaluation and to propose a methodological framework. This framework is not perfect but it is a meaningful beginning. It needs to be developed and supplemented by urban climate modelers of insight in the future.

[Comment 1] Section 2.2. The procedure is relatively well described, but the authors do not discuss the criteria for considering a given PSS value acceptable. This I would imagine would vary depending on the quantity of interest, time, and spatial scales of the problem under consideration.

Response: Thank you for your reminder. We agree with you that the PSS may change significantly by the quantity of interest, time, and spatial scales in different problems of interest, and so generating a reliable standard of 'acceptable PSS values' cannot be fully dependent on one single study - it has to be a joint effort over time. This study was intended to make a first step in this effort, and the standard will likely improve as more researchers apply the PSS method to many quantities, time, and spatial scales.

Quantities of Interest

In light of your comment, we checked the variations of PSS values in our study scope due to different quantities and time of day/year. Figure 1 shows the variations of PSS values due to different quantities of interest. We also checked the statistical significance of the between-group difference using the t-test. P-values among the groups of PSS values for different quantities show that no significant ($p<0.05$) difference was found among T-2, ST2, RH, and W10, while ST14 and Precip had significantly ($p<0.05$) lower PSS values compared to the other attributes but the difference between the average levels of the largest and the smallest group was below 0.2. Therefore, it is possible to have a unified standard of acceptable PSS values while highlighting the standard can be relaxed slightly for specific quantities known to have lower reliability.

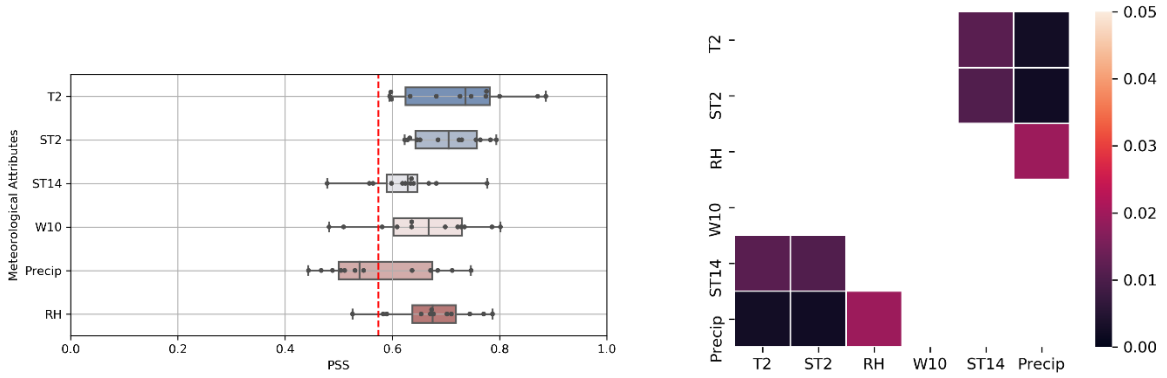

Figure 1 Variations in the PSS values due to different quantities (left) and the t-tests among the PSS values for different quantities (right). The red dashed line indicates the 75% quantile level among all PSS values.

Time

We also checked the variations of PSS values over the time of year. PSS values in all months of the year had mean/median PSS values larger than the 75% threshold we proposed. No statistically significant ($p<0.05$) were observed among any monthly groups of PSS values. It is for future studies to check further how PSS values change over time, for example, ten years ago or later, with significant changes in meteorological contexts.

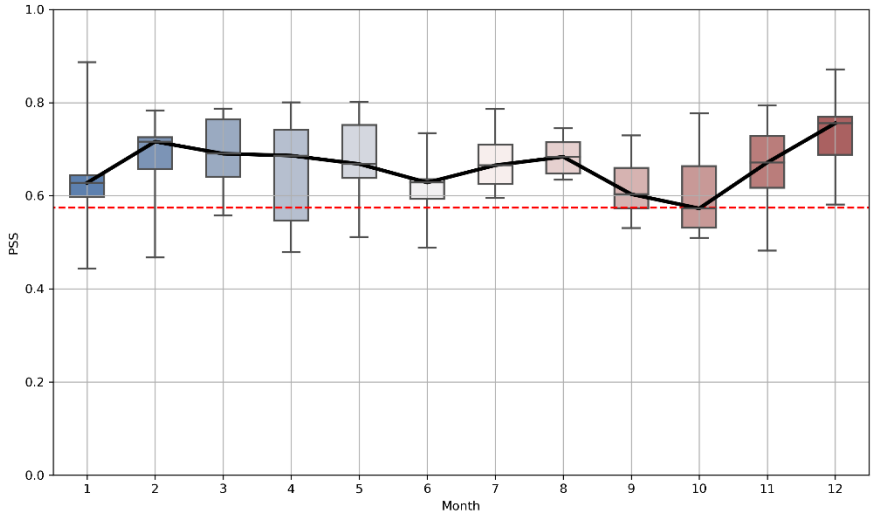

Figure 2 Variation of PSS values in different months of the year.

Spatial scales

Since all simulations conducted in this paper use the same spatial extent and scale, we cannot thoroughly check how PSS values vary over different extent and scales. However, it is reasonable to claim that the proposed standard of acceptable PSS values in this study sets the minimum requirement since simulation accuracies were usually found higher for simulations having lower spatial resolutions due to the spatial-smoothing effects. Simulations using coarser spatial resolutions should at least meet our standard of acceptable PSS values, and the standard can be tightened in future studies using coarser spatial resolutions.

[Comment 2] English requires substantial revision: Several sentences are qualitative or poorly formulated, and several typos are present throughout.
Response: Thank you for the nice reminding. We did our best to correct these errors.

[Minor comment 1] P1L30. ". . . urban climate simulation models are among the most powerful ones." –> This sentence is not very accurate. What do the authors mean by "most powerful"?
Response: Thanks for your comment. We have changed 'powerful' to 'widely-used.'

[Minor comment 2] P1L34. "its corresponding observed ones." –> "and corresponding observations."
Response: Thank you very much for the reminder. We have made revisions accordingly.

[Minor comment 3] P1L35-37. "Model" or "modeling is used eight times; please rephrase avoiding repetitions.
Response: Thank you very much for the reminder. We have made revisions accordingly.

[Minor comment 4] P2L1-2. This sentence is a repetition of the concepts explained in the preceding paragraph. I suggest removing it or rephrasing.
Response: Thank you very much for the reminder. We rephrased the sentence.

[Minor comment 5] P2L10. "of every conclusion" –> "of conclusions".
Response: Thank you very much for the reminder. We have made revisions accordingly.

[Minor comment 6] P2L35. "interval" –> "departure"?
Response: Thank you very much for the reminder. We have made revisions accordingly.

[Minor comment 7] P3L30. "instinct" –> A rigorous procedure rather than instinct should be adopted to assess whether model results compare well against experimental measurements.
Response: Thanks for your reminder. We have revised the sentence as follows,

"Therefore, we included three different temporal resolutions in our model evaluation framework (Table 1) - annual, monthly, and daily - to provide a sophisticated view on whether the modelled results could replicate the temporal and spatial patterns in the observations or not."

[Minor comment 8] Fig 4. "Comparaison" –> "Comparison" (title of the figures)
Response: Thank you very much for the reminder. We have made revisions accordingly.

[Minor comment 9] P13L28. When comparing point-wise measurements with grid-cell averaged simulation results, some kind of upscaling procedure should also be adopted. Can the author address this problem?
Response: Yes, you are right that there exist interpolation methods available to transform point-based observations to grid-based data analytical results. However, doing so will introduce more uncertainty associated with the interpolation method used and the parameter selected or optimized in the interpolation method. Therefore, in this paper, we chose a more explicit path by directly comparing the observations with the grid-based simulations. We also pointed out the potential risks in this comparison for the readers to consider whether to improve their practice adopting methods such as interpolation or not.

[Minor comment 10] P13L33. "Theoretically, verifying or validating a model is impossible." –> please specify which "model".
Response: The model specified the numerical model in the earth science in the paper of Oreskes et al. 1994 (Oreskes, N., Shrader-Frechette, K., & Belitz, K: Verification, validation, and confirmation of numerical models in the earth sciences, Science, 263(5147), 641-646, DOI: 10.1126/science.263.5147.641, 1994.). The atmospheric model also is a numerical model in earth science. The model specifies the atmospheric model in our manuscript. We added a sentence as follow at the beginning of Subsection 5.1 and revised the "model" to "atmospheric model":

"*The atmospheric model also is one of the earth-scientific numerical models.*"

Response to Reviewer 5

[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-to-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]

(1) The authors touch one of the most important problems of model development and using, the model verification workflow. I completely agree with authors that is very important research question, which frequently is often omitted in the model-based research. This scientific problem is even more important for urban climate modelling due to the lack of the urban-scale observations and high complexity of urban climate processes.

Response: Thank you for your comments.We are apprectated that you deemed the topic of this study impontant.

(2) However, in my opinion, the manuscript is far away from being accepted. The biggest problem of the manuscript is briefly described in the next few sentences. Starting just from the manuscript's tile, authors refer to the problems of the urban climate research and modelling. However, the presented results poorly fit typical urban climate research framework and do not touch the known problems of urban climate modelling. The urban climatology & meteorology typically works with anomalies such as urban heat island, urban dry/moist islands, urban-induced precipitation anomalies, etc. The state-of-the art

mesoscale models, such as WRF, COSMO or HIRLAM, are still not perfect in terms of accurate simulation of these anomalies. The development of the common evaluation & verification methodology for urban climate is a very relevant research question. However, the presented results deals practically nothing with indicated research problem. The presented results could not tell the reader, how good or bad is the considered model in terms of the simulation of the specific urban climate features. During the previous stages of revision, the authors have provided results focused on urban-rural differences (e.g. Figure 5). However, these figures are still useless for the evaluation of the urban climate modelling, because the observed and modelled values are spaced apart in different subplots.

Response: Thank you for your comments. However, we think that the reviewer mixes up two different topics: simulate urban climate features and model evaluation. Urban climate features' simulations are the research topics to seek out new urban climate features by using modelling technologies. The trustworthiness of the results of these research needs to be established by model evaluation. The model evaluation is the comparisons of the modeled meteorological variables with its corresponding observed ones. In this study, we only have five meteorological observed data: air temperature, MODIS surface temperature, 10-m wind speed, 2-m relative humidity, and precipitation, and accordingly we only can conduct the comparison between these observed data with its corresponding modelled ones. The purpose of this paper is to tell the readers the importance of model evaluation which has been overlooked by previous urban climate research and proposed a methodological framework of model evaluation which has been mentioned in previous literature.

(3) The presented results raise many questions even in isolation from the specific problems of urban climate modelling. For example, the authors provide a great amount of the same type of graphs with a seasonal variation of the observed and modelled values. The manuscript and supplementary materials are strongly overloaded by similar graphs. What do they want to show by this plenty of graphs?

Response: The model evaluation is the comparisons between modelled variables and its corresponding observed ones. Each modelled variable has a set of graphs. The type of graphs is same, but the variable is different. Therefore, it is necessary to shows all comparisons between modelled variables and its corresponding ones even if the graphics are similar.

(4) It is trivial that regional climate model, forced by the realistic reanalysis data, could reasonably simulate the seasonal cycle of the key weather variables. More interesting and relevant question is how the high-resolution mesoscale model captures the regional climate features.

Response: The reviewer mixes up the model evaluation and the research findings in urban climatological features. The findings in urban climatological features retrieved from the urban climate modelling results usually cannot evaluate directly. Model evaluation is a critical step of quality assurance to the modelling results. Therefore, model evaluation is a responsibility of climate modellers for establishing the trustworthiness to the findings, which is the reason why we emphasized the importance of the model evaluation in urban climate researches.

(5) The unique dense observational network could provide a lot of information on this topic. However, this part of analysis is omitted in the study. Even the questions related to the diurnal cycle are more relevant due to the well-known biases of the daytime and nighttime biases of the models. But this type of analysis is given much less attention in comparison to the analysis of the seasonal variations.

Response: In fact, we provided the comparisons in the daytime and night-time variations between each observed meteorological variable and its corresponding modelled ones.

(6) Finally, it is important to show the advantages of the proposed verification framework and statistical scores. The authors criticize the simpler approaches of the model verification in the introduction. But what benefits do the presented approach give in comparison to the simpler approaches (e.g. the simple biases or model-to-observation plots, which are frequently used in model-based studies)? It would be amazing to present, e.g., that presented framework allows to identify some model errors that could not be revealed by simpler methods. But this is missing. The authors only present the model scores for different variables. But with what these values should be compared?

Response: Thank you very much for your comments. It is a good future research direction.

(7) I suggest that the manuscript should be significantly revised before acceptation to GMD. In my opinion, one way of revision is adding more focus to the regional climate features and, specifically, urban climate features such as urban heat island and its quantitative metrics. Otherwise, the authors should not claim about urban climate modelling in the title, abstract and introduction.

Response: We don't agree the comments. The model evaluation is different with the researches in regional climate features, such as urban heat island, urban dry/moist islands, urban-induced precipitation anomalies, etc. The model evaluation is a quality assurance to the new findings in regional climate features' researches. The paper intend to remind urban climate researchers the importance of model evaluation and provide a methodological

framework for it.

(8) In addition, the are some other specific issues related, which are listed below. Please, note that these comments are addition to the general comment, but not its detailed explanation. In other words, resolving only the indicated specific questions is not sufficient for the revision.

Response: The paper had been revised many times. We are confident that it is valuable to publish in high impact journal before it reveals a pain-point which the urban climate modellers never paid enough attention on the model evaluation.

[Specific Comment 1] P1, L30-31: The discussed tools and models are very different in terms of scale and complexity. It will be good to prove some examples of the certain tools and models.

Response: It is not the focus of this paper to discuss the difference of tools and models in terms of scale and complexity. This sentence was changed as follows:

In this vein, many tools have been developed, and the rapidly developing urban climate simulation models are among the most widely-used ones.

[Specific Comment 2] P1, L33 – P2, L8. I completely agree with general idea of the paragraph. However, there is a number of studies, where detailed verification of urban-scale models is performed. This studies should be indicated in the literature review together with studies, where verification part is omitted or not sufficient.

Response: We added more details of previous studies which model evaluation part is omitted or not sufficient in Section S10 of Supplementary Material.

[Specific Comment 3] P3, L1-21, sect 2.1 (and also P6, L1-9). It is common to present more detailed information about the model setup in such regional modelling studies. E.g. the scheme of the nested domains is missing. For the urban climate modelling studies, it is common to present more detailed information about the study area and land use/land cover data. What are the exact list of the urban land cover parameters, used by the model? What are the typical values of the urban faction and anthropogenic heat flux in the study area? How the used parameters were obtained? These information is required to comparing of the presented study to other urban climate modelling studies. Referring to the dataset name, even without a literature reference, in insufficient.

And more general remark: I suggest to join the two indicated sections to one section, related to urban climate model and its setup.

Response: We provided the information of the urban climate model and its setup in Sections S1 and S4 of Supplementary Material. Moreover, we will provided all detail information in other two papers (*A high-resolution urban land surface dataset to investigate the urbanization impact using urban climate modelling: 1979-2010* and *Quality assurance in high-resolution urban climate simulation: using WRF ARW/LSM/SLUCM model (version 3.7.1) as a case study*).

[Specific Comment 4] P3, L18-21: Is 1-day spin up sufficient for development of the urban climate features in the model?

Response: The 1-day spin up is enough to each 4-days simulation segment.

[Specific Comment 5] P8 (Figure 3). Why do you plot modelled values with so low temporal resolution (only 4 values for a day, the same temporal resolution as it is in FNL reanalysis)? The key advantage of the high resolution regional climate model is opportunity to increase the resolution of the driving gridded meteorological data (FNL reanalysis in our case), both in space and in time. Using the model output with 1-hour resolution with be improve the results and the presentation quality.

Response: The urban climate simulation is a computational resources consuming job, especially computing wall time and storage space. In this study, one 4-days simulation segment need 1 day node-computing wall time and 28 G storage space. The temporal span of urban climate simulation in this study is one year. Total 122 4-days simulation segment need 122 days node-computing wall time and 3 T G storage space. The simulation job outputs data per hour would be need huge computational resources. Moreover, this study is just for presenting a methodological framework, and accordingly 6-hour temporal resolution is enough.

[Specific Comment 6] P8-9 (Figure 4, 5): As I noticed in the general comment, these figures seems to be useless for model evaluation, because the modelled and observed values are displaced to different subplots.

Response: We don't agree. For the interpretations of Figures 4 and 5, please refer to Pg6,Ln30 to Pg7, Ln2.

[Specific Comment 7] P10, L13-15: Please, clarify, good in comparison to what? E.g. you could compare the presented score by the score, obtained for original gridded data (FNL reanalysis), or by the score obtained by WRF model with different, or with some scores from literature. I

have the same comment to the places in the text where same scores for other variables are presented and discussed.

Response: The PSS may change significantly by the quantity of interest, time, and spatial scales in different problems of interest, and so generating a reliable standard of 'acceptable PSS values' cannot be fully dependent on one single study - it has to be a joint effort over time. This study was intended to make a first step in this effort, and the standard will likely improve as more researchers apply the PSS method to many quantities, time, and spatial scales. Moreover, we already explained the reasons why the modelled variable differ from its observed ones in Subsections 5.2.

Other comment to these lines: as I've noticed in general comment, it is trivial that model captures the seasonal cycle, and that the seasonal cycle is more-o-less the urban and rural areas. More attention should be addressed to diurnal cycle and spatial variations within the study area.

Response: In fact, we provided the comparisons in the daytime and night-time variations between each observed meteorological variable and its corresponding modelled ones.

[Specific Comment 8] P11, L21-23: Please, clarify, acceptable for what? As I've noticed from the Supplementary materials, the mean model bias for the surface temperature could be quite high, up to 10K. Why do you consider it acceptable? The same comment is for the next section related to wind.

Response: The PSS may change significantly by the quantity of interest, time, and spatial scales in different problems of interest, and so generating a reliable standard of 'acceptable PSS values' cannot be fully dependent on one single study - it has to be a joint effort over time. This study was intended to make a first step in this effort, and the standard will likely improve as more researchers apply the PSS method to many quantities, time, and spatial scales. Moreover, we already explained the reasons why the modelled variable differ from its observed ones in Subsections 5.2.