

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.

Major Comments:

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:
  - a) **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).
  - b) **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.
  - c) **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.
  - d) **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  $\pm 3$  degree in temperature change should have higher impact than  $\pm 3$  millimeter in precipitation. The authors should consider how to choose reasonable intervals for different variables.
  - e) **Selection of variables:** the authors should state the rationale for choosing variables for model evaluation in your case study.
  - f) **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.

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.
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.
  - a) 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?
  - b) Schemes of physics components: How to choose the schemes for each component? Would the selection of schemes have impacts on PSS scores?