Comments on the revised manuscript "Model evaluation of high-resolution urban climate simulations: using WRF/Noah LSM/SLUCM model (Version 3.7.1) as a case study"

General comment: 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.

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.

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? 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. 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.

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?

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.

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.

Other specific comments:

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.

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.

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.

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

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.

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.

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.

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.

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.