Response to the Referee

We thank the Referee for the insightful and constructive comments on our manuscript. The format is as follows:

Referee's comment
Our reply (revised text)

Referee #1

General comments

The manuscript titled "A new Earth Observation–based WRF configuration for urban regional climate simulations over Paris" aims to evaluate the added value of improved Urban Canopy Parameters (UCP) in the Weather Research Forecast model (WRF) when using an Urban Canopy Model (UCM). The study focuses of the city of Paris (France) for a short time period in the summer of 2020 when a significant heatwave occured.

The study compares 3 different WRF experiments. In the first, the default urban parametrization of WRF is selected, where artificial surfaces are simulated by the land surface model following a "bulk" approach. In the other two experiments, the BEP-BEM multi-layer UCM is used with two different UCP datasets: one based on a LCZ map and default parameters, and another with updated UCP adapted for the Paris region using on Earth Observation data.

This study is very interesting, but could benefit from some improvements regarding the manuscript's structure, as well as a more in-depth analysis of the differences observed between the experiments. I have provided various comments to this effect, which I hope will improve the manuscript's readability, and the strength of its conclusions.

We thank the Referee for the feedback. The manuscript has been revised in terms of its structure, following Referee's recommendations. New figures have been also added to the supplementary material. Further sections and figures have been added regarding the evaluation of the PAR experiment, to examine thoroughly its divergent behavior in non-urban areas compared to the other two simulations.

Because of the substantial revisions made to the original manuscript, we recommend that the reviewers assess the revised manuscript using the version in which all tracked changes are accepted. Nonetheless, a version displaying the tracked changes will be also provided.

Accordingly, all responses to the reviewers' comments refer to the clean revised manuscript version, in which the track changes have been accepted.

More generally, I felt that the study would have benefited from a more in-depth analysis or more detailed hypotheses when examining the results regarding the difference between the WRF experiments and the potential added value of the improved UCP, as it seemed a bit too descriptive as it stands. For example, one could combine an element similar to Figure 6 with a specific examination of how UCP parameters evolve at the precise location associated with the station, and see if it is possible to make sense of this evolution (for example, if a station sees its urban fraction increase between PAR and CTRL, what is observed in terms of the temperature response).

We agree. For this reason, we added in the manuscript a table (Table 3) with the percentage differences in the derived UCPs, between PAR and CTRL, as well as two figures (Figures S1 and S2) in the supplementary material, whereas a schematic representation of the morphological UCPs (building and road widths and height distribution) between the two simulation is presented, as a means to quantitatively show the morphological differences for each LCZ class between the PAR and CTRL.

Specific comments

Structure of the manuscript

Regarding the structure of the manuscript and its various sections, the distinction between urban cover parameters derived from Earth observation data, presented in subsection 2.2.1, and the more general section 3 devoted to land-use data did not seem clear to me.

I understand that you are moving from WRF "general physics" to urban parameterizations, and then to the required urban parameters, but Section 2.3 then seems inconsistent, as you have just introduced a high-resolution dataset specific to your study area and then return to more general datasets on land cover. Furthermore, in the current structure, you have only a single sub-sub-section 2.2.1.

One suggestion might be to clearly separate, on the one hand, the physical aspects and urban parameterizations, and on the other, the land surface datasets and parameters. It might look something like this:

2 Data & Methodology

2.1 The Weather Research and Forecasting Model

2.1.1. Model configuration and physics

2.1.2. Urban parametrizations

2.2 Land Use Forcing

2.2.1. Standard land use data in WRF

2.2.2. Improved urban canopy parameters derived from EO data

The restruction of section 2 has been been made following Referee's recommendations.

Another general point, which I return to several times in my subsequent comments, concerns the difference between point-based analyses and masks. I feel that it is always difficult to combine these two approaches, because it is extremely unlikely that the masks we construct will match the characteristics of the stations used for the evaluation. One way to get around this problem could be to split the analysis into (1) a purely point-based evaluation (vs. station), which focuses more specifically on the description of the station and the UCP used (even if only for a few stations), and then (2) a more general comparison of the experiments (without the stations) that can take masks into account: for example, the urban and rural masks you have defined, but it might also be interesting to group the model points based on their LCZ.

The point-based analysis aimed to evaluate the performance of the three simulations against observations, while the two masks were made to evaluate each simulation in the representation of the urban heat island and its intensity as a phenomenon. The inclusion of the stations inside the two masks aims to a qualitative comparison of the modeled and observed urban heat island, following a same approach made in Langendijk et al., 2025 (Langendijk, G. S., Fernandez, J., Demuzere, M., Diez-Sierra, J., Quintana, Y., Fita, L., Zazulie, N., Nogherotto, R., Carril, A. F., Chun, K. P., Giuliani, G., Halenka, T., Hoffmann, P., Muñoz, L. E., Pietikäinen, J.-P., Rechid, D., and Yuan, J.: Representation of global mega-cities and their urban heat island in CORDEX-CORE regional climate model simulations, npj Urban Sustainability, https://doi.org/10.1038/s42949-025-00325-6, 2025.)

For clarification purposes, we added this statement in section 3.4:

"We also calculate the observed UHII comparing the mean temperature of the stations lying within the urban mask to that of the stations in the rural mask (black dashed line) as a qualitative comparison (Langendijk et al., 2025)."

Section 2 Data & Methodology

I think it would be helpful to add a few sentences introducing Noah-MP (beyond simply mentioning it in Table 1) in subsection 2.1. That way, when you mention the various land surface models used with WRF later (line 92), readers will already be familiar with them. Furthermore, I think it might be interesting to introduce the dominant coverage approach used by WRF. In this regard, I was wondering if you had considered using the "mosaic" approach introduced by Li et al. (2013), now that you are moving to increasingly precise land surface datasets? [After rereading the whole section, I'm wondering if you're referring this in line 98? See my comment below].

Li, D., Bou-Zeid, E., Barlage, M., Chen, F., & Smith, J. A. (2013). Development and evaluation of a mosaic approach in the WRF-Noah framework. Journal of Geophysical Research: Atmospheres, 118(21), 11-918.

We added some further information about Noah-MP in section 2.1.1 (section 2.1 in the older version of the manuscript). Furthermore, we found that the line 92 comment ("In this approach, which is integrated inside WRF's any given land surface model (LSM)..") wasn't in fact correct, since in README.namelist file inside the run folder of WRF, it is clarified that sf_urban_physics option and the corresponding urban canopy models of this option are only coupled with Noah and Noah-MP land surface models.

Regarding your comment about the mosaic approach, this can be implemented in Noah LSM, while we used Noah-MP LSM. Noah-MP LSM divides each grid cell into two tiles. Below is the statement added in section 2.1.1 covering the abovementioned issues:

"The Noah-MP (Niu et al., 2011; Yang et al., 2011) land surface model (LSM) has been utilized, since, compared to its conventional version Noah LSM (Tewari et al., 2004), it improves the representation of land–surface heat and water exchanges across a range of surface types, including snow–covered terrain, grasslands and forests (Barlage et al., 2015; Chen et al., 2014, 2016; Gao et al., 2015). Furthermore, Noah-MP LSM applies a division of each grid cell in two tiles (vegetated and bare soil fractions), and the corresponding biogeophysical and biochemical processes are calculated separately for each."

About the line 98 comment ("This coupling involves activating UCMs for the urban fraction within urban grid cells only, replacing the land surface model for this fraction when computing fluxes (e.g. heat and momentum)"), this in fact is happening inside subroutine noahmp_urban located in module_sf_noahmpdrv.F, a module which is at a folder named phys in WRF chain.

Line 78. "Convection is explicitly resolved in the inner domain (d02)"

I was wondering if you used the same parent domain simulation (d01) to run the three different urban experiments presented in Section 2.4, and, if so, which urban parameterization is enabled at the 12 km resolution? I assume the BULK approach is being used? I'm not certain, but it might be worth mentioning somewhere – either here or in Section 2.4 – in a brief sentence when you present the experiments.

Yes, we used the same parent domain in terms of coverage and all physics options except for urban physics option. Thus, in BULK d01 had sf_urban_physics=0, while in CTRL and PAR d01 had sf_urban_physics=3. We added a statement in section 2.3 (former section 2.4) clarifying the configuration in each simulation:

"Hence, the simulations conducted are: 1) PAR, in which urban canopy parameters specifically tailored to the Paris region and derived from EO data were applied; 2) CTRL, with the default UCPs used in WRF version 4.6.0; and 3) BULK, in which no UCM is activated for the parent and nested domain. BEP–BEM UCM was implemented in PAR and CTRL for both domains, while all three simulations had the same parent domain (d01) in terms of coverage and all the rest of physics options, as depicted in Table 1."

Line 98. "This coupling involves activating UCMs for the urban fraction within urban grid cells only, replacing the land surface model for this fraction when computing fluxes (e.g. heat and momentum). The LSMs compute these fluxes for the remaining non–urban fraction. The final fluxes are area–weighted and averaged for the entire cell."

If I understand this sentence correctly, you are indeed using the mosaic approach implemented by Li et al. (2013)? If so, this should be clearly stated, as it is not always the case in WRF studies (for example, the simulations conducted for the FPS-Convection). Following up on that, my question is whether this approach is used consistently across your various simulations; for example, what happens in the BULK experiment? Finally, it might be interesting to present maps of the urban fraction, for example in additional panels in Figure 3.

The mosaic approach is only available in Noah LSM (sf_surface_physics=2) through namelist parameter sf_surface_mosaic, while in all the experiments we implemented Noah-MP (sf_surface_physics=4). A statement about what happens in BULK has been added in section 2.1.2:

"In this approach, there is no fractional split between vegetated and bare soil for the urban grid cells, since the vegetated tile has zero weight. Consequently, all grid cells classified as urban are assigned generic urban properties (e.g. roughness length, surface albedo, heat capacity, thermal conductivity) through category ISURBAN in MPTABLE.TBL, without reference to the specific characteristics of a simulated city."

Regarding the urban fraction for the two BEP-BEM simulations (PAR and CTRL), it was found that urban fraction values were not taken into account from the lookup table URBPARM_LCZ.TBL but from the wrfinputs. As a result, both experiments had the same gridded urban fraction values (not per LCZ class) through the WUDAPT to WRF (W2W) process (Demuzere, M., Argüeso, D., Zonato, A., and Kittner, J.: W2W: A Python package that injects WUDAPT's Local Climate Zone information in WRF, J. Open Source Softw., 7, 4432, https://doi.org/10.21105/joss.04432, 2022). This process has been utilized in the framework of FPS URB-RCC for the creation of the geo_em files. A sentence has been added to section 2.3 regarding this issue:

"PAR and CTRL shared the same values for the rest of the UCPs and urban fraction was defined on each grid cell by the W2W process."

Line 109. "In this way, the vertical distribution of sources and sinks of heat, moisture and momentum induced by buildings is explicitly represented, which shapes the thermodynamic profile of the urban roughness sub–layer and, consequently, the urban boundary layer, in more detail than SLUCM"

Another point that I could see discussed is the difference in which the two UCMs are coupled to the atmospheric model. Kusakas's SLUCM is only coupled to the atmosphere at one level located above the canyons rooftops – at the top of the urban roughness sublayer – while BEP can be coupled at multiple levels, depending on the number and height of the atmospheric models closest to the ground.

Thank you for your comment. We revised this issue with this sentence in section 2.1.2:

"The scheme neglects the variation in building height within each urban land use category and treats the urban canopy as a single homogenous layer located below the first atmospheric layer of the mesoscale WRF model (Hang et al., 2024). On the other had BEP (sf_urban_physics=2) parameterizes a three–dimensional urban morphology within a multi–layer framework, enabling variations in building heights and building density, defined by the road and building widths in urban grid cells (Martilli et al., 2002). In this way, the urban canopy is resolved in multiple vertical layers, which lie within the WRF atmospheric column but do not necessarily coincide with the model's vertical levels. Consequently, this multi–layer treatment in BEP enables the vertical distribution of sources and sinks of heat, moisture and momentum induced by buildings to be explicitly represented, which shapes the thermodynamic profile of the urban roughness sub–layer and, consequently, the urban boundary layer, in more detail than SLUCM."

Line 120. "2) thermal and 3) human related parameters."

I was wondering if it might be helpful to briefly mention that the LCZ classification was designed to group different urban areas based on their morphology and is therefore closely linked to building height, building density, etc., but is less closely linked to physical parameters that depend more on building materials. For example, even though a neighbourhood of a similar LCZ in Paris (France) and Shanghai (China) may have buildings of equivalent height, the materials used are likely to differ, which affects the physical properties of the walls, roofs, etc. This could potentially be raised later to justify the need to for improved UCP using EO data.

We agree with your comment. LCZ classification is heavily based on the morphology of the buildings. Table 4 in Stewart and Oke, 2012 includes values of thermal, radiative, and metabolic properties for local climate zones and it is mentioned that "values given here are therefore subjective and should be used cautiously". We added the following statement in section 2.1.2:

"In short, this classification system defines 17 land cover types, of which 10 are urban (built type) classes, primarily based on morphological characteristics such as building heights (high–rise, mid–rise, low–rise) and building density (dense, open, sparse)."

Line 130. "Starting from WRF v4.6.0, an updated lookup table with more representative values is introduced as default in WRF".

At first, I thought these updated values came from EO data for Paris. After rereading the text, if I understand correctly, these are the new "default" values to be used with BEP and the LCZs? Is there a study we can refer to regarding this update, or could this work serve as a reference in the future? If so, would it be advisable to also include the old values in Tables S1 and S2 for future reference? That might make the tables too complicated or too cluttered, so I understand why you might disagree.

Furthermore, you note that the old default values may not be very realistic, particularly for European cities. While the new values may be better suited to Europe, could they be less suitable for other regions of the world as a consequence?

Correct, these are the new "default" values to be used in WRF's URBPARM_LCZ.TBL lookup table and were based on fixes made by Alberto Martilli (CIEMAT, Spain). You can find the corresponding pull request on github here. The main issue in the previous "default" lookup table was the large values for road width and thermal properties for specific LCZs (you can find the corresponding WRF pull request on github here. In this pull request, it is mentioned that road widths for some LCZs were set to 108 meters, resulting in occasional model crashes and unrealistic surface temperatures of > 350 K. Regarding your last question, this is something that we are not the experts to answer, since the modifications were made by Martilli. Nevertheless, our understanding is that the previous road width values of 108 meters are unrealistic for any given region in the world, so the new "default" lookup is more suitable than the previous one, regardless of the area of interest.

Line 140. "Additionally, UCPs related to anthropogenic urban heat emissions remained unchanged from the default values WRF provides and were assumed to be constant over 24 hours."

I suppose this refers to an additional flux directly added in the model? If so, are there different values between the LCZ? If not maybe you could provide the value for the BULK-class?

Answer: Yes, anthropogenic heat emissions have different values among LCZs and are based on a variety of parameters. You can find these parameters in the URBPARM_LCZ_Paris.TBL lookup table that we used for the PAR simulation (https://zenodo.org/records/19253784). Since these parameters were not changed, you can assume that for CTRL simulation they remain the same for each LCZ. More precisely, the parameters affecting anthropogenic heat emissions are:

TIME_ON
TIME_OFF
TARGTEMP
GAPTEMP
COP

TARGHUM
GAPHUM
PERFLO
HSEQUIP
HSEQUIP_SCALE_FACTOR
SW_COND
BETA

Bulk simulation uses parameters already defined in MPTABLE.TBL in Urban category 13 (“Urban and Built-up”), which does not include parameters for anthropogenic heat emissions.

Line 146. “2.2.1 Urban canopy parameters derived from EO data”

Would it be possible to include a table listing all the Earth observation data used, their sources, their original horizontal resolution, and the urban canopy parameters they help improve? This seems to be an important part of the article.

Details about the UCPs derived from EO data have been added in section 2.2.2.

Line 163. “Land Use Data”.

As suggested earlier, this section could be divided into two subsections, one for the “standard” WRF configuration and one for the updated data based on EO.

After rereading this subsection several times, I’m no longer sure I fully understand what’s being done. You mention the LANDMATE PFT dataset as well as the IGBP-MODIS classification, then a modified IGBP-MODIS map, and finally the WRF-LCZ map.

Here is what I understand so far:

(1) You start with the LANDMATE PFT dataset to follow the EURO-CORDEX simulation protocol

(2) You "modify" its urban class to replace it with the default IGBP-MODIS urban class in order to obtain the appropriate parameters for the BULK simulation; but do you retain the LANDMATE "location" of the urban points?

(3) For the BEP simulation, do you use a modified IGBP MODIS that replaces the urban class with the LCZ? But this is also the case for the BULK simulation outside urban areas: the non-urban areas are identical in both experiments in Figure 3. So, where was LANDMATE ultimately used?

I would suggest adding an introductory sentence specifying which default land cover map/classification WRF uses, followed by a sentence explaining what you modify in your "default" BULK simulation, and then how you improve it using the LCZ map.

Thank you for pointing out this ambiguity. We modified section 2.2.1 so that the workflow of the land use data would be more clear. As a result the workflow is as follows:

1) LANDMATE PFT dataset was used and then translated to the International Geosphere–Biosphere Programme – Moderate Resolution Image Spectroradiometer (IGBP MODIS) classification inside WRF. As a consequence, for BULK simulation, LANDMATE "location" of the urban points has been retained.
2) To add the LCZ classification in the land use for BEP-BEM experiments, WUDAPT to WRF (W2W) tool has been utilized, leading to a replacement of all urban grid cells (which were represented until now by one land use category, category 13 "Urban and Built-up") from BULK and a part of non-urban grid cells (due to the tool) by the LCZ classes.

Line 180. "it was found that urban grid cells in the BEP–BEM simulations (LCZ classes) cover a larger portion of Île-de-France (23.4%) compared to the urban area represented in the BULK simulations (16.6%)"

Following up on the previous comment, I would emphasize that this is due to the land cover map associated with each experiment, and not the experiments themselves (which is what you're doing by mentioning the LCZ classes for the BEP simulation). You could very well have chosen to define the LCZ map as a common "default" and simply replace each urban LCZ class with the IGBP-MODIS urban class for the BULK experiment.

Your interpretation is indeed correct. We could have used the modified land use from W2W process to BULK, since we found that all LCZ grid cells, when namelist parameter sf_urban_physics is 0, all LCZ classes fall back to the one urban category 13 "Urban and Built-up". We weren't aware of this when we conducted the simulations. Thus, our study is primarily a sensitivity study of the urban parameterizations, however, when comparing the BEP-BEM simulations with the BULK simulation it also contains a land use sensitivity element over the specific grid cells that changed from/to the "urban" category.

After reading this section, I'm once again wondering about the "fraction/tile" approach mentioned earlier (see my previous comment regarding line 98). How does WRF handle the remaining non-urban fraction of urban points? I assume it involves using NOAH, but where does it obtain the information on the required natural land cover and the associated parameters?

This comment has likely been already answered from previous statements.

Line 189. "2.4 WRF simulations"

I understand why you're presenting this section next to the "Observations" section, but for now, it focuses mainly on the urban and rural masks definition, and very little on the experiments. This is tricky since you've already presented the WRF configuration in Section 2.1 and the Earth observation data in Section 2.2.1. I would suggest adding a sentence explaining the choice to simulate this specific period, perhaps by mentioning the design of the FPS-URB-RCC experiment, which aimed to simulate the 2020 heatwave in Paris.

Thank you for this comment, we added in section 2.3 the following statement:

"In this study, three simulations have been conducted under the umbrella of STAGE-0 simulations of the FPS URB-RCC protocol, which focuses on the examination of RCMs, urban schemes and the overall model and domain configuration. Simulations of STAGE-0 were planned to be centered over the Paris region, covering a period that includes two extreme weather events in 2020, namely a convective event during May and a heatwave during August, thus facilitating a coordinated analysis of both extremes. Here we focus on the latter event and the broader summer period of 2020, for the simulations to be assessed under both extreme and natural conditions, evaluating their ability to accurately reproduce extreme events and their overall long-term behavior."

Line 205. "Observations"

I would suggest, if possible, mapping all the stations used (33?) and not just those located in the urban and rural masks. I just noticed that they appear in Figure 3; you could probably remove them from there, since they are neither presented nor discussed.

One idea might be to include two maps in Figure 4: one showing all stations and their corresponding LCZ colors, and a second map – the one you already have – highlighting the stations used for the urban heat island analysis.

We have added a figure (Figure 4) in the revised manuscript with the location of all the stations used for the evaluation of the three experiments.

Perhaps a table containing some information about the stations would be helpful (for example, their names, coordinates, elevation, whether they are located in a park, etc.). Not for all of them, but perhaps for the urban ones, or those located in the masks?

We have added in the supplementary material Table S5, which contains information for all the stations used.

More generally, in this section, what is the rationale for using different stations for the temperature analysis and for the UHI analysis, and how can the two be reconciled?

Temperature analysis takes place over the corresponding grid cells of the stations and we chose to include all the available stations to produce a more robust result. We are referring to the 33 stations that fulfilled the 1% missing value criterion and that did not have land use discrepancies between their RADOME classification and the model.

The UHI analysis relies on the masking process as a more elaborate and strict distinction between the urban and rural grid points, in order to better isolate the impact of the urban environment. Thus, we chose to include only the stations that lay within the aforementioned mask as better representatives of the urban and rural parts of the mask. Thus, the "observed" UHI from the stations is better matching the "modelled" UHI since they both are refined through the same masking process.

Line 208. "To minimize the impact of data gaps, only stations with less than 1% missing observations relative to the total number of expected records were retained"

What do you mean by the total number of expected records? Wouldn't the period under analysis be sufficient? I would also suggest mentioning it again: from May to August 2020.

RADOME network provided with 51 stations with data (observations) per 6 minutes, of which some stations had missing data for a large amount of the period. So we decided to use for evaluation only those stations with less than 1% missing observations. A quantitative example for a day would be that a station's maximum number of observations (records) are 240, so if a station has 230 records, then it is excluded from evaluation.

Line 209. “station classifications”.

What do you mean by that? Is it simply the urban/rural distinction based on your urban and rural masks? But then what happens to the station outside of the rural mask?

The distinction of the stations to urban and rural are based on a urban/rural classification already given by RADOME network, according to the GHS-UCDB polygons delimiting urban areas. To investigate whether the station types indicated by RADOME network match the land-use categories represented in the simulations, it was found that 9 stations do not match the real land use due to the resolution of the model grid. The correspondence between stations and the model was made by assigning to each station the nearest model grid cell, using a nearest-neighbor approach.

Line 212. “Consequently, a total of 33 stations”.

You could mention in the previous sentences how many stations were excluded due to missing data and how many because of a mismatch in the land cover.

We assessed section 2.4 and pointed out the number of stations excluded due to a mismatch of land cover, while mentioning the initial number of stations provided by the RADOME network.

After reading this section and the previous one (2.4), I understand that you use all the points from the urban and rural masks for your UHI analysis (Figure 10), using the averages of these masks as references. However, no explanation is provided regarding the method used for your analysis of urban and rural temperatures. It is only in the results section that you explain that, for the LCZ sub-analysis, you extract the model point closest to each station. My question, therefore, is why you rely on these two masks for your UHI analysis, especially since you are comparing it to stations.

I understand that the method by Diez-Sierra et al. (2025) was designed to create urban and rural masks in a standardized/automated manner, but I am not convinced that it is appropriate for evaluating a model when using stations; there is very little chance that the model masks are representative of the stations (or vice versa).

We believe this issue has already been answered earlier, nonetheless the use of the two masks has the goal to visualize and evaluate the intensity of the UHI and how the phenomenon is represented in each experiment. Figure S5 has been further added for this purpose. The use of the stations in Figure 14 is added as quantitative comparison, similarly to the approach made in Langendijk et al., 2025.

Section 3 Results

Line 223. "nighttime (18:00–07:00 UTC) than during daytime (06:00–17:00 UTC)".

I think the distinction between day and night is fine as it is, but for additional information, I would like to point out that in Paris in June (as an example, when the days are longest) the sun rises around 06:00 UTC+2 and sets around 22:00 UTC+2. Another way to distinguish between day and night could also be to use the downwelling shortwave radiation from the model.

Thank you for your comment, indeed ideally we could divide nighttime and daytime hours based downwelling shortwave radiation. We believe that through the daily profiles figures the performance of the three simulations can be seen, despite a possible discretization of daytime and nighttime for the statistical evaluation.

Line 230. "In non–urban areas (Fig. 5b)".

The fact that the results of the BULK and CTRL experiments are extremely similar makes sense, since nothing was changed in the model at those specific points. These slight differences could be explained by the effect of changes made at other points (urban), which have a limited impact on neighboring points through advection in the atmospheric model. But then, how can we explain these significant differences in the PAR experiment? According to Figure 3, it appears that the modified IGBP-MODIS map shows a different urban extent; are we certain that the model points classified as rural are represented in the same way in the PAR simulation as in the others?

We thank you for noticing this. In fact, it was something we assessed by adding a distinct new section in the manuscript (section 3.3). As a result, our findings point to a possible internal model variability for the explanation that PAR is significantly colder and divergent than the other two simulations. We also added this reference: Bassett, R., Young, P. J., Blair, G. S., Samreen, F., and Simm, W.: A Large Ensemble Approach to Quantifying Internal Model Variability Within the WRF Numerical Model, Journal of Geophysical Research: Atmospheres, 125, https://doi.org/10.1029/2019JD031286, 2020.

Before delving into the LCZ analysis, I noticed something in Figure S1: in the observations, all temperatures (urban, rural, daytime, and nighttime) increase as the summer progresses – with August > July > June – but this is not always the case in WRF for urban temperatures. The CTRL and BULK experiments appear to show *higher daytime peaks in July than in August. This likely points to the heat wave that occurs in August and WRF's ability to simulate it.*

Your finding is true about BULK, it seems that for urban stations the daytime peaks in August are lower than in July. The following statement regarding this has been added in section 3.1:

"Month to month variation is shown in the supplementary material (Fig. S3) and it is noticed that, while observational temperatures increase from June to August, BULK and PAR during midday hours in urban stations estimate bigger temperatures in July than in August."

Line 245. "Although the BULK simulation does not use the LCZ land–use classification to urban grid cells (unlike the two BEP–BEM simulations), we extracted BULK values from the corresponding grid cells so that diurnal cycles could be compared across LCZ classes (Fig. 6)."

This sentence gives the impression that you are only extracting the points from the model that are closest to the stations for this LCZ analysis, and that this had not already been done for the temperature analysis in Figure 5. But it was already done for Figure 5, right?

In Figure 5 (now Figure 6 in the revised manuscript version), for each simulation, we took all 13 stations that have their corresponding modelled grid cells an urban related land use type ("Urban and Built-Up" for BULK, LCZ 2, LCZ 5, LCZ 6 and LCZ 8 for BEP-BEM runs) and grouped them all in a category called urban.

Then, in an effort to also examine the impact of the discretization that the LCZ classification provides for urban areas, an additional grouping was made based on the LCZ class of each urban station in the BEP-BEM simulations. Of the 13 urban stations, three were located in grid cells classified as LCZ 2, three as LCZ 5, five as LCZ 6 and two as LCZ 8. Essentially, in Figure 6 (now Figure 7 in the revised manuscript version), we extracted BULK values from the corresponding grid cells of each LCZ group, in a way to point the discretization the LCZ classes offer compared to the one urban class BULK uses.

Line 258. "Furthermore, BULK shows warm biases of similar magnitude across all LCZs, which reflects the limited urban variability captured when urban areas are represented using a single land–use category."

I understand that you mean that, since all urban points have the same parameters in the BULK experiment, we would expect to observe similar urban temperatures. But does that necessarily imply similar biases? Indeed, if the stations are truly representative of the LCZ in which they are located, we would expect their temperatures to differ; consequently, if the BULK experiment does not reflect this variability, the biases should not be identical.

Your comment is indeed correct, and we revised this sentence with the following in section 3.1:

"Furthermore, despite the limited urban variability when urban areas are represented using a single land–use category BULK shows warm biases of similar magnitude across all LCZs and has a worse performance than PAR and CTRL for the rest of the LCZ stations."

Line 259. "The analysis separated into daytime and nighttime period (Table S5) indicate the same pattern found in the urban stations group, with substantially warmer biases during the night than midday for both BEP–BEM simulations across all LCZs."

Do you have any idea why this is the case? At first glance, one might expect the UCM to improve results at night, since they should better accounts for urban heat storage than the simpler BULK model.

It is reported that PBL scheme used (Boulac) is responsible for the warm biases during the night. the BouLac Planetary Boundary Layer scheme is found to cause nighttime overestimations of temperature in urban related grid cells (Segura et al., 2021), thus also partially explaining the nighttime warm bias not only of BEP–BEM simulations but also that of BULK. A future work might include a sensitivity study with the added value of each PBL scheme.

Line 266. Table 3.

One suggestion I would make is to also add the BULK indicators to the table. Another idea that comes to mind is this: would it make sense to also include the observation values so we can get an idea of the temperatures observed across the different LCZs? But that would probably require reorganizing the table first by LCZ, then Observation | BULK | CTRL | PAR, to avoid repeating the observation row… But that might no longer match the text. It's just an idea; I understand if you don't go with it.

We added BULK indicator in in Tables 4 and 6 (Tables 3 and 5 in the older version of the manuscript).

Line 270. "We investigate the model's performance..."

That first sentence mentioning the precise date of the heatwave with a link to the Copernicus website could be moved to the "WRF simulations" subsection.

We added the sentence in section 2.3 and modified it as follows:

"The heatwave event occurred from 7 to 13 August 2020 and affected western Europe, including the Paris region, with temperatures reaching up to 40 °C (https://climate.copernicus.eu/surface-air-temperature-august-2020)."

Line 310. "Figure 9 presents the averaged 2 m air temperature differences"

I was wondering if it might be helpful to include the PAR temperature maps as well, in addition to the delta you've shown. If you want to keep the maps the same size, you could reorganize the figure by placing the nighttime maps in a left column with (1) PAR temperature, (2) the PAR-CTRL difference, and (3) the PAR-BULK difference; and do the same in a right column for the daytime maps.

We added PAR temperature values in the left column of the new figure (Figure 10 in the newest version of the manuscript).

Line 316. "The colder daytime climatology of PAR at the non–urban areas is also apparent spatially when compared with the CTRL and BULK simulations (Fig. 9c,d)."

I touched on this briefly earlier when discussing the differences observed in the experiment regarding rural temperatures, but since this is even more striking here, I find myself asking the same question: do we have any idea what might explain such marked discrepancies outside urban areas? Could it be that the variations observed in urban areas located within the small domain, but outside the Paris region, have an impact on regional-scale flows? Because an average drop of 2 °C (figure 9.d) over the duration of the heat wave seems quite significant.

As mentioned earlier, we added a whole section (3.3) regarding PAR behavior in non-urban areas.

Line 330. "The urban mask contains 109 grid cells."

This part made me wonder whether you used the same masks for all 3 experiments and, if so, what dataset is it based on, because, if I understand correctly, it uses certain urban characteristics, such as the urban fraction?

We mention in section 2.5 that the urban and rural masks were based on threshold values of variables inside WRF, such as the urban fraction and additional static inputs such as terrain height and land–use fraction. This procedure effectively excludes the influence of altitude on temperature (i.e., adiabatic lapse rate), enabling a more accurate assessment of the UHI. This is based on the methodology of Diez-Sierra et al. (2025), by using different threshold values.

Line 358. "Figure 10".

I could see at least two other things that could be discussed from Figure 10. The first is that the spread of the UHII (shown in pink) provides insight into the added value of the UCM compared to the BULK approach, as the latter exhibits consistently lower *spread because all urban points share identical surface characteristics. Similarly, the differences in spread between UHII in CTRL and PAR are most likely related to differences in UCP, but it is difficult to determine which ones, and which experiment is the best. The second point concerns the weak temporal correlation between the three experiments and the observations. In this regard, I was wondering if you had directly looked at a similar plot for urban and rural temperatures compared to the observations. It might be interesting to see if WRF is capable of capturing the slow development of the heat wave outside the city and how urban areas subsequently react.*

We agree with your first finding, therefore we added this statement in section 3.4:

"All simulations are able to represent the nocturnal UHI. In BULK, grid cells within the urban mask show a smaller temperature range than the BEP–BEM simulations, highlighting the lack of urban variability in this approach. Deviations from the averaged UHII (thick red line) are larger in BEP–BEM simulations than in BULK, indicating a stronger model response to the urban heterogeneity introduced by the LCZ classification."

Regarding your second finding, we added a new figure in the manuscript (Figure 11), in a way to examine also the temporal agreement between the three simulations and the observations during the heatwave, for urban and non-urban stations.

Section 4 Summary and Conclusions

Since this section summarizes the manuscript and because of all the comments I have made previously, I will wait for the authors' response before commenting here.

Technical corrections

Line 52. "Currently, the Coupled Model Intercomparison Project (CMIP) protocols represent urban environments in a homogeneous manner, with limited differentiation among cities."

I even wonder if the CMIP global climate models actually represent cities at all, given their coarse horizontal resolution.

Your comment is indeed correct. Even in CMIP6 HighResMIP models cities may not be represented as a grid cell, rather than a subgrid fraction or urban tile. We modified this sentence as follows:

"Currently, the Coupled Model Intercomparison Project (CMIP) protocols hardly represent urban environments in a homogeneous manner, due to their coarse resolution."

Line 52. "the the"

Typing error.

We corrected the typing error.

Line 64. "EO data".

The acronym could be introduced here for the first time (outside the abstract).

This has been settled.

Line 66. *Add (France) next to Paris.*

This has been settled.

Line 81. "All simulations are driven by the ECMWF ERA5 reanalysis dataset (Hersbach et al., 2020) at a horizontal resolution of 0.25° × 0.25° and 6–hourly temporal intervals."

You could mention the use or absence of spectral nudging, following the CORDEX and FPS protocols.

We didn't use spectral nudging, therefore we modified this sentence as follows:

"All simulations are driven by the ECMWF ERA5 reanalysis dataset (Hersbach et al., 2020) at a horizontal resolution of 0.25° × 0.25° and 6–hourly temporal intervals, without the use of spectral nudging."

Line 92. "i.e. urban parameterization switched off".

This is a somewhat semantic question, but the BULK approach in WRF could still be considered a form of urban parameterization, since some climate models do not model urban areas at all: some replace them with vegetation cover, while others replace them with "rocks" without any adjustment to surface parameters.

This has been an ambiguous issue. We noticed in ICUC12 last year that people from United States consider BULK an urban parameterization, while Europeans do not. Nonetheless, we added this statement in section 2.1.2:

"Although no UCM is activated, BULK could still be considered a form of an urban parameterization, since the land use cover is adjusted by the LSM, in a way that parameters are modified for urban grid cells to better account for the zero-order effects of urban surfaces, particularly with respect to wind speed and heat storage capacity (Chen et al., 2011; Liu et al., 2006)."

Line 92. "which is integrated inside WRF's any given land surface model (LSM)".

I would suggest rephrasing the sentence as follows: "which is incorporated into all the various land surface models (LSMs) available in WRF."

As mentioned earlier, we found that BULK approach – through namelist parameter sf_urban_physics=0 – can be implemented only in Noah and Noah-MP LSMs.

Line 92. "(e.g. roughness length, surface albedo, heat capacity, thermal conductivity)".

It might be helpful to include the values for each of these parameters; that way, readers can use them for comparison when you later present the old and new UCM parameters in Table S2.

These values are hardcoded and can be found in the technical documentation of Noah-MP.

Line 99. "replacing the land surface model for this fraction".

You can replace "land surface model" with LSM.

This has been settled.

Line 107. "BEP (sf_urban_physics=2), on the other hand, parameterizes".

I would suggest reversing the order of "BEP" and "on the other hand" to make the text easier to read: "On the other hand, BEP (sf_urban_physics=2) parameterizes..."

This has been settled.

Line 130. "Starting from WRF v4.6.0, an updated lookup table with more representative values is introduced as default in WRF."

I was wondering if there are any references regarding these changes, and perhaps a study analyzing the sensitivity of the new values?

The issue about the unrealistic values in the older "default" version of the lookup tables URBPARM_LCZ.TBL has been addressed in github. We have added this statement in section 2.1.2, clarifying where this has been issued:

"However, it was found that these default values for some parameters affecting urban morphology (e.g. road widths for some LCZs) and thermal properties were unreasonable and can become unrealistic for European cities, given the heterogeneity of European urban structures. This issue was introduced through WRF Pull Request no. 1954 (https://github.com/wrf-model/WRF/issues/1954) and settled in WRF Pull Request no. 1969 (https://github.com/wrf-model/WRF/pull/1969) and incorporated in commit 0a62e59. As a result, an updated LCZ lookup table with more representative values was introduced and implemented for the FPS URB-RCC STAGE-0 simulations which, starting from WRF v4.6.0, it was defined as the default lookup table in WRF."

Line 189. "WRF simulations".

I think the title works, but I was trying to come up with another one to distinguish it more clearly from the title of subsection 2.1, "WRF model configuration and physics." Something that would better highlight the fact that this section describes various sensitivity experiments involving urban canopy parameters.

We believe that after the reconstruction of the sections in the paper and the appropriate modifications, section WRF simulations describes better the experiments than before.

Line 214. "("Urban and Built–Up" for BULK, LCZ2, LCZ5, LCZ6 and LCZ8 for BEP–BEM runs)".

Could change "," for ";" between BULK and LCZ2 to separate the lists.

This has been settled.

Line 232. "1.62°C"

Typing error, missing a space before °C. Coming back to this comment, you should double check the same typo happens in other places.

This has been settled.

Line 264. Figure 6.

You could add the LCZ name in the panels title as you did in the text: LCZ2 (compact high–rise).

If we understand correctly, you suggest adding the "compact high-rise" part in the panels title. If this is the case, we believe the LCZ classes can still be separated as is without the description of each LCZ class.

Line 270. "during a heatwave event occurred"

Should add "that" or "which" between "event" and "occurred".

This sentence has been moved to the section 2.3 and it has this form:

"The heatwave event occurred from 7 to 13 August 2020 and affected western Europe, including the Paris region, with temperatures reaching up to 40 °C (https://climate.copernicus.eu/surface-air-temperature-august-2020)."

Line 332. "indicating the range of UHII variability"

maybe add "indicating the range of UHII variability relative to the average rural temperature".

This has been settled.

We again thank the Referee for their thorough comments and evaluation. We believe that the revisions have enhanced the consistency of the manuscript.