Reply on RC2 'Comment on gmd-2023-122'

We are pleased that you are in favour of our manuscript and we would like to thank you for your suggestions for some corrections.

Comment 1:

"*Ln. 61*

The authors state that 'Nevertheless, the emulated ML model can never exceed the accuracy of the numerical model because it is trained based on the model's results', but in the results and discussion it turns out that the proposed ML models often give better results than numerical models. I propose to explain this contradiction."

Response 1:

Indeed. In some situations, the ML models are more accurate than the numerical models. There are several reasons for this. First, the UTCI estimated by the ML model is closer to the observations because it combines several downscaling models (T_{mrt}, T_a, RH, and wind speed) that take urban form and function into account. SOLWEIG, on the other hand, does not downscale T_a, RH, and wind speed. Another reason is simply luck. An ML model with high accuracy compared to a numerical model sometime has the error on the 'right side', which may lead to slightly higher accuracy when compared to observations. Nevertheless, we added a clarification to the manuscript (Line 344):

The lower RMSE of the HTC-NN compared to SOLWEIG can be explained by the combination of four submodels that downscale T_a, RH, T_{mrt}, and U separately, while SOLWEIG downscales only T_{mrt} comprehensively.

Comment 2:

"*Ln. 72*

'… four cardinal wind directions …' – I am used to analyzing the wind field as three-dimensional. Did I misunderstand something?"

Response 2:

Thank you for this comment. We have calculated the wind speed from the x, y, and z components. The four cardinal wind directions only describe the general inflow direction of the LES and ML models. That is, we computed statistical wind fields using x, y, z components, but only for four wind directions due to computational cost. We added the following sentence at line 72:

The wind fields are calculated from the x, y, and z wind components.

Comment 3:

"Ln. 213

'The error distributions of SUEWS and the MLP across the different stations are similar (Fig. 3a)…' – I think that Fig. 3a shows the error for all stations rather than the error distributions across stations."

Response 3:

You are right. Fig. 3 does not show error distributions of the models but boxplots. The boxplot shows all errors of all stations. We changed this in the text (line 221):

The errors ~~distributions~~ of SUEWS and the MLP across the different stations are similar (Fig. 3a), with a higher variability during the night than the during day.

Comment 4:

"*Ln. 227*

'The T_{mrt} U-Net has a slightly lower accuracy than SOLWEIG (RMSE of 6.18 K to 5.86 K; R2 of 0.84 to 0.86)' (and also in ln. 239) – the acceptable level of accuracy is usually a subjective choice. However, for T_{mrt} in the standard ISO7726 (ISO, 1998), ISO recommends that the error in T_{mrt} estimates should be within ±5°C. Could you please address/discuss this."

Response 4:

You are right. ISO7726 recommends that measurements or models should estimate T_{mrt} within ±5°C. On average we are indeed within ±5°C, as the mean absolute errors (MAE) are 4.25 K and 3.83 K for the ML model and SOLWEIG respectively. However, the root mean square error is higher as this error metric gives more weight to outliers. Due to complex shadow patterns it is very difficult to always predict T_{mrt} accurately, even for numerical models (e.g., see also Fig. 4 (b) and Briegel et al. 2022). It is inevitable to always model T_{mrt} within the ±5°C range. Therefore, for the purpose of modelling outdoor urban thermal comfort, we believe that the achieved model accuracy of the ML is sufficient.

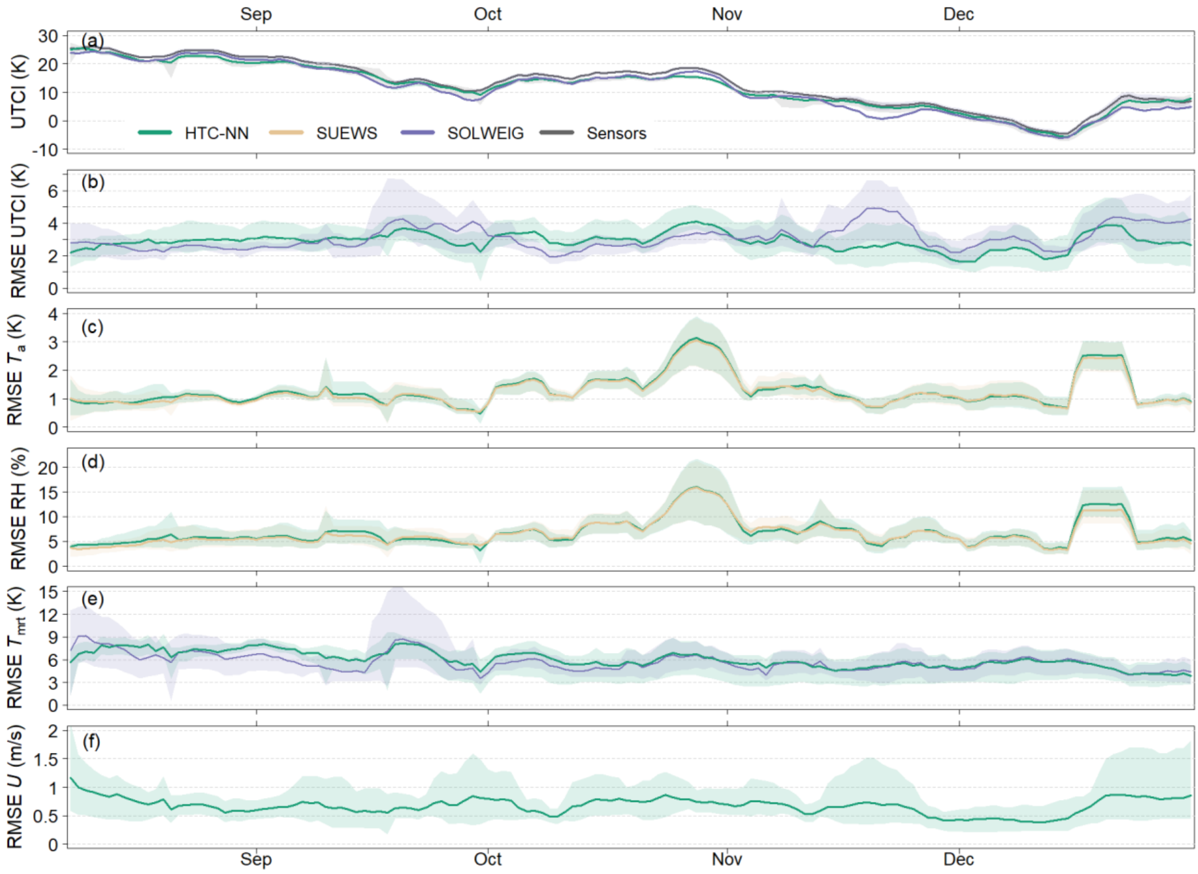
Comment 5:

"*Fig. 4*

The RMSE of SUEWS predictions (orange lines) are almost invisible - could they be bolded? Please change "Dez" to "Dec"."

Response 5:

Thank you for this suggestion. We changed it. The orange lines showing the SUEWS results are still not perfect to see. That is because the errors of SUEWS and the ML model are almost in line which makes it hard to make the both error lines perfectly visible.



Comment 6:

"Fig. 6

Some of the sharp drops in T_{mrt} and UTCI in the SOLWEIG charts (e.g. afternoon 2022-02-11 at the station marked "e") are likely the results of shading, which are directly calculated by SOLWEIG at ground level, while the reference data is from 3.5 m. Similarly at early morning or afternoon at other stations. Am I right? Anyway, could you comment on these rapid, sawtooth changes in the SOLWEIG charts and their effect on the accuracy statistics."

Response 6:

This is actually a very good point. The complex shadow patterns within the city as well as the different heights between model and observations could be reasons for the rapid and strong changes in the T_{mrt} and UTCI predictions. We have not investigated the impact of these strong outliers on the overall accuracy of the model. However, your question is also related to comment 4: RMSE are higher than MAE by 1-2 K. The reason for this could be the strong outliers you mention in this question. The MAE of SOLWEIG ranges from 1.99 to 3.30 K and the MAE of HTC-NN (U-Net) from 1.74 to 3.28 K for the different sensor stations. The plot you mention (Fig. 6 (e)) has actually the highest overall MAE of all stations with 3.30 K, while Fig. 6 (b) has the lowest with 1.99 K. Fig. 6 (b) does not show any rapid, sawtooth changes.

So, we can conclude that a more detailed investigation of the error patterns would be beneficial for further research and model development. But as we don't intend to evaluate numerical models in this study this would exceed the scope of this research.