General comment:

The manuscript addresses the need for accurate thermal comfort mapping in urban environments, which increasingly experience heat waves. Here, the authors present the approach they have developed: the Human Thermal Comfort Neural Network (HTC-NN). The modeling approach demonstrates a comprehensive integration of numerical models and machine learning techniques to predict human thermal comfort in high resolution urban environments. Evaluation is done for Freiburg, Germany in predicting street-level Universal Thermal Climate Index (UTCI) with high spatial resolution using street-level measurements. Variations in thermal comfort and hot spot disparities during different times of the day are identified at the neighborhood-level scale.

Overall, the manuscript provides a well-structured overview of the study. I have enjoyed reviewing it and I congratulate the authors with this work: it makes for an impressive contribution to the field of urban climate research; however I do have some remarks for improvement mainly in the evaluation section. I recommend it is accepted with minor revisions as detailed below.

Specific comments:

Section 2.4

Here it would be nice to have some absolute statistics about the reference period, to provide some context. What is mean JJA temperature (or annual cycle) over the study area, quantify the heat extremes in this period (how many, what temperature?), etc from ERA5.

Section 2.5

The method is quite complex. It is a very important part of this manuscript as this is a development/technical paper. There can be more structure in the modeling approach section, and in the section of each submodel's development.

The way I understand it now is the following: The development of HTC-NN involves four main steps: First, initial spatial and meteorological data are generated from various sources. Then, so-called "ground truth" data for the four HTC-NN submodels (MLPs, U-Net, RF) is calculated using numerical models. This includes data for air temperature (Ta), relative humidity (RH), mean radiant temperature (Tmrt), and wind speed (U). The third step is the training and evaluation of the submodels, which include two Multi-Layer Perceptrons (MLPs) for modeling Ta and RH at the neighborhood scale, U-Net for modeling Tmrt at the building-resolved scale (that has already been trained and validated in an earlier study) and a Random Forest (RF) model for statistical wind fields. The final step involves linking the submodels and calculating the Universal Thermal Climate Index (UTCI).

In particular I suggest the authors to make this section more explicit as to what parts are physically modelled, and what parts are trained submodels. It is physical modelling part of the HTC-NN? Line 141 is confusing to me; I am expecting 3 subsections in 2.5 (i.e., two MLPs and RF), but there are 4. Further, I think it will be beneficial to explicitly state per submodel what it takes as input and forcing data. Regarding the physical modelling/preprocessing; consider adding an extra subsection to section 2 where you can explain how you have used LES and SUEWS.

Section 2.6

Please expand the explanation of UTCI, elaborate and provide the definition of the heat stress classification groups. Later in your analysis you use these terms: strong, very strong, or extreme heat stress (e.g., l.279).

Section 3.1 and 3.2

Looking at figure 1, your sensor data is mainly situated in urban sites, while your model area has a considerable fraction of more open fields. That may skew your observations. Please validate the Ta, RH, and U submodel components as well as the UTCI temperature with ERA5 and/or other types of reanalysis data full training period (2018-2022) . That will clarify whether the errors you find, such as the peaks in October and December (l.220), are robust.

Section 4

In the discussion section, the authors thoroughly discuss the model's performance and limitations. It is highlighted that HTC-NN demonstrates a favorable balance between computational cost and accuracy compared to numerical models. Remarkable errors are discussed, among which diurnal error patterns impacting UTCI predictions and the model's tendency to overestimate UTCI during the day during heatwaves. Limitations include the absence of coupling with a mesoscale model, neglecting local weather phenomena, and individual modeling of Ta and RH tiles. Dependency on data from a single weather station and potential initial error propagation are acknowledged issues. Suggestions for improvement involve coupling with mesoscale models and using reanalysis data.

Can you elaborate more generally on the limit of NNs, particularly training a network with a limited amount of extreme events?

Figure 1:

I find the gray grid cells in the figure are not well visible. Perhaps you can experiment with a different shade of gray, or explicitly mention in the caption something along the lines of "Note: Gray grid cells indicating the training areas of the Ta and RH submodels may be less visible due to color contrast."

Table 1 and 2:

Please write out the used abbreviations (LCC, DEM etc.) at their first use.