Responses to Reviewer #2's comments:

This study evaluates HRRRv4 forecasts against two observational networks in Lubbock, Texas: the dedicated U-HEAT deployed across the city, and the regional West Texas Mesonet. The U-HEAT dataset is a clear strength of the paper and provides a valuable basis and a detailed year-long assessment of systematic model biases. The inclusion of nocturnal cooling rates and urban heat advection in the evaluation is an important contribution as it extends the analysis beyond standard meteorological variables. The manuscript is well organised, with clear sections, and is relevant for both urban climate studies and operational forecasting applications. At the same time, certain aspects of the study could be clarified and extended to further strengthen the generalisability and reproducibility:

Thank you for your valuable feedback. Please see our point-to-point responses below.

Major comments:
1. Since the study is centred on a single mid-sized city in a semi-arid climate, it would strengthen the conclusions to discuss more explicitly how the identified biases might generalise to other small cities under different climatic conditions. A brief paragraph clarifying transferability across different climatic regimes would help readers gauge generalisability.

Thank you for your suggestion. We will add a new paragraph to clarify the transferability, and the last two paragraphs of the revised Conclusions will read:

"Although this evaluation focuses on a single small city in a semi-arid climate, several of the identified forecast biases are likely to occur in other small cities under different climatic conditions. This expectation arises primarily from HRRR's use of a slab urban scheme, which simplifies urban surfaces, and is partially supported by previous evaluations of near-surface temperature and wind speed at non-urban sites. However, confirming the transferability of these biases will require dense, city-scale observational networks deployed in additional small cities. This is particularly important because many small urban areas are represented by only a few HRRR urban grid cells, yet can exhibit substantial spatial variability in vegetation fraction, soil moisture, and urban morphology.

Future work should advance evaluation and model development in parallel. Replicating this analysis in other small cities with similarly dense within-city observations will enable more systematic assessments of model performance across different climatic regimes. From a modeling perspective, our findings underscore the need for more realistic urban representations in NWP systems. Future developments should prioritize the integration of advanced urban canopy parameterizations, refined sub-grid land surface heterogeneity, and high-resolution urban observations. More broadly, this evaluation highlights the limitations of applying conventional NWP systems to urban environments without targeted enhancements. As cities face growing challenges from extreme heat and flooding, poor air quality, and evolving land cover, integrating urban-specific processes into NWP frameworks (Wang et al., 2025) will be essential to ensure accurate, actionable forecasts in both research and operational contexts."

Reference:

Wang, C., Zhao, Y., Li, Q., Wang, Z., and Fan, J.: Ultrafine - Resolution Urban Climate Modeling: Resolving Processes Across Scales, J. Adv. Model. Earth Syst., 17, e2025MS005053, https://doi.org/10.1029/2025MS005053, 2025.

2. The evaluation of nocturnal cooling rates (Sect. 3.3) is informative, but is based on a subset of nights with continuous domain-wide cloud cover below 25% and statistically significant cooling (p<0.05) (Sect. 2.5). To assess robustness, it would help to report how many nights satisfy the cloud-cover filter and to briefly justify the chosen threshold at 25%.

Thank you for this helpful comment. Applying both filters, i.e., domain-mean cloud cover below 25% and statistically significant cooling ($p < 0.05$), yields 41 nights. If only the statistical significance criterion is applied, 51 nights are retained. The 25% cloud-cover threshold follows the U.S. National Weather Service definition (https://www.weather.gov/bmx/nwsterms), where 12.5–25% cloud cover corresponds to "mostly clear or mostly sunny" conditions. As a sensitivity test, using a stricter 12.5% cutoff (i.e., clear or sunny) results in 40 nights, with no change in the conclusions. We will add these counts, the rationale for the threshold, and a note on sensitivity to the manuscript:

"… we restrict our analysis to nights with continuous domain-wide cloud cover below 25%. This threshold, corresponding to "mostly clear" conditions in U.S. National Weather Service definitions, is selected to isolate surface-driven cooling processes and minimize cloud-related variability while retaining an adequate sample size."

"Note that we also evaluated the sensitivity of the results to the selection criteria. Using only the statistical significance criterion ($p < 0.05$) yields 51 nights, whereas applying a stricter 12.5% cloud-cover threshold results in 40 nights. The conclusions remain unchanged across these sensitivity tests."

Minor comment:
The acronym UHA is introduced in the abstract but not defined at its first occurrence in the main text (Sect. 1, line 97). Please ensure that acronyms are consistently defined when first used in the manuscript.

Thank you for pointing this out. We will add the full name of UHA at its first occurrence in the revised Introduction section.