I hope that the following comments will be helpful for further improving the authors' work.

The revised manuscript has been substantially improved compared with the original submission, and many of the revisions described in the Response letter have been incorporated into the revised manuscript. In particular, the correction of the normalized mean absolute error (MAE) values to inverse-transformed degrees Celsius values, the introduction of symmetric mean absolute percentage error (sMAPE), the use of out-of-fold stacking, the Bayesian Ridge Regression (BRR) formulation, Multi-Input Multi-Output (MIMO) forecasting, Shapley Additive Explanations (SHAP) analysis, input-configuration comparison, and multi-site validation are substantial improvements.

However, at this stage, several important issues still require further clarification and revision. Some concern inconsistencies between the explanations given in the Response letter and the descriptions in the revised manuscript, while others concern internal inconsistencies within the revised manuscript itself. In particular, the leakage-free claim, resampling artifacts, the MIMO implementation, the stacking and BRR meta-learner inputs, the scope of uncertainty quantification, the interpretation of solar-radiation proxies, and the claims regarding methodological novelty require further attention.

I list these points below as major and minor comments.

Major comments

1. The authors present the Improved LSTMs Ensemble with Stacking (ILES) framework as a leakage-free stacking ensemble. The revised manuscript describes out-of-fold base-learner predictions for training the BRR meta-learner, which addresses leakage at the stacking level.

However, the preprocessing description in the revised manuscript, lines 317-324, states that short missing gaps are filled by linear interpolation between adjacent valid observations. Unless this interpolation is performed in a forecast-time-aware manner, it can use observations later than the forecast issue time. This is particularly important because historical road surface temperature is used as an input while road surface temperature is also the prediction target. For example, imputing a missing historical road surface temperature value at time t by using the observed value at time t+1 would contaminate a one-hour forecast issued at time t.

Thus, the revised manuscript demonstrates leakage-free meta-learner training conditional on already preprocessed data, but it does not demonstrate that the full preprocessing and forecasting pipeline is leakage-free. The authors should clarify the order of quality control, imputation, hourly aggregation, train-test splitting, and fold construction. They should confirm that no validation or test input value was imputed using observations later than the forecast issue time, and that imputed road surface temperature values were not used as evaluation targets.

Relevant locations: Response letter, response to Major Comment 1, Methodological Work Stream 4; Response letter, response to Major Comment 5; revised manuscript, lines 20-26, 120-124, 232-265, 317-324, and 414-420.

2. The authors' response to Minor Comment 2 in the Response letter states that resampling artifacts were avoided by aggregating precipitation as hourly totals, computing all other variables as hourly means, and treating hours with fewer than six valid five-minute records as missing. In the revised manuscript, however, this procedure is described only briefly: the raw observations were collected every five minutes and then resampled to hourly resolution; precipitation was aggregated as hourly totals; and all other variables were computed as hourly means "to suppress transient sub-hourly fluctuations." The revised manuscript does not quantify how much smoothing this operation introduces into the road surface temperature series, nor does it evaluate whether artificial signal components are introduced by the downsampling procedure.

This issue is important for interpreting the claimed methodological gains. In Table 4, the MAE improvement from configuration 3 to the full ILES framework is 0.383 degrees Celsius to 0.373 degrees Celsius, that is, only 0.010 degrees Celsius. In Table 5, the one-hour MAE improvement of ILES over KNN-LSTM is 0.389 degrees Celsius to 0.373 degrees Celsius, that is, only 0.016 degrees Celsius. If the smoothing effects or downsampling-induced artifacts are larger than these incremental gains, it becomes difficult to determine whether the reported improvement reflects a robust methodological advantage of the proposed architecture or the model's ability to exploit an artificially smoothed or transformed target series.

In particular, when a five-minute road surface temperature series is downsampled to hourly resolution, high-frequency components that cannot be represented at hourly sampling may be aliased into lower-frequency components if they are not sufficiently removed. This can create apparent low-frequency structure or phase shifts that were not present in the original five-minute series. The authors should use the original five-minute road surface temperature series and the hourly resampled road surface temperature series to quantify at least the magnitude of smoothing and the presence or absence of aliasing-induced artifacts, and should show that these effects are negligible relative to the reported MAE values and the incremental gains of ILES over the strongest baseline.

Relevant locations: Response letter, response to Minor Comment 2; revised manuscript, lines 317-324, 473-482, and 520-545; Tables 4 and 5.

3. The authors' response to Minor Comment 5 in the Response letter states that the revised manuscript adopts a MIMO strategy for simultaneous one-, three-, and six-hour forecasting. The revised manuscript adds a MIMO section at lines 289-301. However, the KNN-LSTM and BiLSTM-MHA formulations still appear to define scalar outputs, and the training description states that BiLSTM-MHA uses a "single-unit dense output layer" in the revised manuscript, lines 439-443.

This is not consistent with a multi-output MIMO formulation. The authors should clarify whether the base learners output a three-dimensional forecast vector, such as the forecasts for t+1, t+3, and t+6, or whether separate horizon-specific direct models are trained. The output-layer dimensionality, loss function, and BRR meta-learner formulation should be made consistent with the chosen multi-horizon strategy.

Relevant locations: Response letter, response to Minor Comment 5; revised manuscript, lines 164-187, 216-220, 289-301, and 439-443.

4. The authors' response to Major Comment 6 in the Response letter explains that direct solar radiation is unavailable and that radiation-driven diurnal forcing is indirectly encoded through historical road surface temperature and air temperature. The revised manuscript similarly discusses historical road surface temperature as an input that reflects the current thermal state and notes that the 24-hour window aligns with the diurnal solar cycle.

This proxy argument should be substantially tempered. Lagged road surface temperature and air temperature can encode the historical thermal response at the specific sensor location, but they do not quantify incoming shortwave radiation, local shading, sky-view factor, screening by roadside objects, road orientation, or other microscale exposure conditions that can strongly affect road surface temperature over short distances. Similar air temperature values can correspond to substantially different road surface temperature values under different radiation and shading environments. Therefore, the present model should be described as a station-specific time-series predictor at a fixed exposure, not as a model for nearby shaded or screened road segments.

Moreover, historical road surface temperature is included in the station-only input set but does not appear in the SHAP plots, which show only air temperature, relative humidity, wind speed, and precipitation. A high SHAP importance for air temperature demonstrates a statistical air temperature-road surface temperature association; it does not demonstrate that solar-radiation forcing is adequately represented. The authors should either include historical road surface temperature in the SHAP analysis or clarify that the SHAP analysis is restricted to meteorological variables only. The absence of direct radiation measurements and local exposure descriptors, such as sky-view factor, shading, and road orientation, should be stated as a limitation.

Relevant locations: Response letter, response to Major Comment 6; revised manuscript, lines 325-333, 386-390, 407-410, 560-564, 649-683, and 745-747.

5. The revised manuscript presents BRR as providing "probabilistic forecasts with closed-form uncertainty quantification" and positions this as one of the methodological contributions of the ILES framework. Section 2.4 formulates the BRR posterior predictive distribution, and Section 4.4 reports prediction intervals and empirical coverage. This is a substantial improvement over the original manuscript.

However, the authors' response to Major Comment 5 in the Response letter clarifies that the two base learners, KNN-LSTM and BiLSTM-MHA, provide deterministic point predictions, and the uncertainty arises from the BRR meta-learner. Therefore, the reported uncertainty is the posterior predictive uncertainty of the final linear stacking layer conditional on deterministic base-learner outputs. It is not uncertainty in the physical road surface temperature generation process.

This distinction is important because the revised manuscript introduces the physical basis of road surface temperature through the surface energy balance, including net radiation, sensible and latent heat fluxes, pavement heat storage, and freezing or melting processes. The current BRR formulation does not model uncertainty in these physical processes, unobserved radiative forcing, pavement thermal properties, or surface moisture and freezing processes. The authors should therefore clearly limit the scope of their uncertainty-related claims and state that the quantified uncertainty is restricted to the BRR meta-learner conditional on deterministic base-model outputs. If uncertainty quantification is retained as a strong domain-specific methodological contribution, this limitation should be explicitly acknowledged.

Relevant locations: Response letter, response to Major Comment 5; revised manuscript, lines 24-26, 44-65, 269-287, and 704-720.

6. The authors' response to Major Comment 2 in the Response letter appropriately states that the paper does not claim to invent fundamentally new machine-learning components, but rather proposes a domain-specific methodological framework. However, the framework mainly combines established components: KNN-LSTM, BiLSTM-MHA, stacking, BRR, and SHAP.

The baseline comparison does not fully establish superiority over the closest prior road surface temperature methods. Table 1 is useful as a literature-positioning table, but it is not a same-dataset comparison. Table 5 provides a same-dataset benchmark, but the baselines are not faithful reimplementations of several closely related road surface temperature models listed in Table 1, such as Bai et al.'s attention-based BiLSTM model, Dai et al.'s GRU-LSTM ensemble, or Zhang et al.'s RF-LSTM model.

Furthermore, the incremental gains are small: Table 4 shows only a 0.010 degrees Celsius MAE improvement from configuration 3 to the full ILES framework, and Table 5 shows only a 0.016 degrees Celsius one-hour MAE improvement over KNN-LSTM. Without repeated runs, confidence intervals, or significance testing, it is difficult to judge whether these gains are robust. The authors should either include the closest prior architectures as baselines under the same protocol or substantially temper claims of methodological superiority and novelty.

Relevant locations: Response letter, response to Major Comment 2; revised manuscript, lines 100-104, 473-482, 488-491, and 520-545; Tables 1, 4, and 5.

Minor comments

1. The revised manuscript states that if the lower bound of the 95% prediction interval falls below 0 degrees Celsius, maintenance crews may initiate precautionary pre-treatment. It also refers to the operational transferability of the probabilistic forecasting component.

This is a potentially useful idea, but the revised manuscript does not evaluate road-maintenance decision metrics such as hit rate, false alarm rate, miss rate, precision, recall, or cost-loss performance. Therefore, the operational claims should be phrased as potential decision-support value rather than as demonstrated operational decision-making improvement.

2. The revised manuscript contains some minor typographical and wording issues, including spelling errors such as "focasting," awkward phrasing, and inconsistencies in figure and table captions. These do not affect the main scientific content but should be corrected during final proofreading.

3. Table 1 is useful as a literature-positioning table for road surface temperature prediction studies. However, the reported accuracies may come from different datasets, input variables, temporal resolutions, and evaluation protocols. Therefore, they should not be interpreted as directly comparable performance values. The authors should add a clear note to Table 1 or the accompanying text stating that the reported accuracies are not directly comparable across studies.

4. Several statements claim that ILES or Variable combination 3 is best “across all horizons and metrics.” However, Table 5 shows that ILES is not uniformly best in sMAPE. For example, at the 3-hour horizon, KNN-LSTM has a lower sMAPE than ILES, and at the 6-hour horizon Gated Recurrent Unit (GRU), Extreme Gradient Boosting (XGBoost), and Random Forest (RF) have lower sMAPE than ILES.

Similarly, MS L595–600 states that Variable combination 3 achieves the lowest errors across all metrics and forecasting horizons, but Table 6 shows that its 6-hour sMAPE is 100.355%, worse than combinations 1 and 2. Table 7 also shows that ILES is not best in sMAPE at M9474.

The authors should revise all claims such as “across all metrics and horizons.” A more accurate statement would be that ILES generally achieves the lowest MAE and root mean squared error (RMSE), while sMAPE is not uniformly best.