Dear Reviewer:

We sincerely thank the reviewer for the thorough and constructive comments. Our manuscript egusphere-2025-3638 entitled *A Hybrid Method for Winter Road Surface Temperature Prediction Using Improved LSTMs and Stacking-Based Ensemble Learning* has been substantially revised to address all major and minor concerns raised. We respond to each comment in full below. All line numbers correspond to the revised manuscript.

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."*

Response:

We thank the Reviewer for this careful and precise comment. The concern has two separable parts: (a) whether imputation of missing RST values in the input series used observations later than the forecast issue time, and (b) whether imputed RST values were used as evaluation targets. Before addressing each in turn, we first clarify the complete order of operations in the preprocessing pipeline, as the Reviewer requests.

(1) Pipeline order of operations.

The preprocessing pipeline for all three stations follows the strict chronological order below:

Step 1. Outlier detection. Physically implausible 5-minute values are flagged and set to missing (RST outside [−40, 50]°C; negative precipitation; RH outside [0, 100]%).

Step 2. Missing value imputation. Short gaps (≤ 1 consecutive hour, i.e. ≤ 12 successive missing 5-minute slots) are filled by linear interpolation between the nearest valid bracketing observations within the same period; long gaps (> 1 consecutive hour) are filled using the climatological hourly mean computed exclusively from the training period, combined with spatiotemporal information from neighboring stations.

Step 3. Chronological train–test split. The 5-minute dataset is partitioned at a fixed calendar boundary (training: December 2020–February 2023; test:December 2023–February 2024). Short-gap interpolation within each period uses only observations from that same period; no observation from the test period enters the training-period imputed series, and vice versa.

Step 4. Hourly aggregation. Within each period independently, valid 5-minute records are aggregated to hourly resolution (precipitation as hourly total; all other variables as hourly means). Hours with fewer than six valid 5-minute records are treated as missing at the hourly level.

Step 5. Fold construction. The training-period hourly dataset is subdivided into three temporal folds, each corresponding to one complete winter season, for the out-of-fold cross-validation used to train the BRR meta-learner (Sect. 2.3).

This ordering ensures that no downstream step can introduce observations from a later time into an earlier period's imputed or aggregated values. Steps 3 and 4 are particularly relevant to the Reviewer's concern: because the train–test split precedes

hourly aggregation, and because each period is aggregated independently, the hourly RST values used as model inputs and evaluation targets in the test period are derived solely from 5-minute observations belonging to that test period.

(2) Overall missing rates across all stations and periods.

Table S1 summarises the fraction of 5-minute RST samples involving any form of imputation at each station, for training and test periods separately.

Table S1. Sample size and loss rate of training and test sets at each station at 5-minute resolution.

Station	Training samples	Training loss rate (%)	Test samples	Test loss rate (%)
M9393	77760	1.93	26208	1.34
M9474	77760	8.00	26208	1.10
M9448	68832	5.10	26208	0.13

The overall imputation rates are low, particularly during the test periods. At M9393 and M9474, more than 98% of test-period 5-minute RST samples are fully observed without any imputation. At M9448, only 0.13% of test samples involve imputation.

(2) Short-gap interpolation and long-gap interpolation.

All reported loss rates above are computed at the native 5-minute resolution. However, the relevant unit of analysis for assessing leakage is the natural clock hour, not the individual 5-minute slot, because all model inputs and targets are constructed from hourly-aggregated values.

For gaps not exceeding a consecutive hour (Short-gap events), linear interpolation is applied using the nearest valid observations bracketing the gap. If the right-side interpolation anchor falls within the same natural hour as the missing slot(s), the resulting hourly mean incorporates no observation from any later hour, and no cross-hour information enters the model by construction. Only when the anchor falls in a later calendar hour than the gap does any forward-looking information enter the corresponding hourly value.

For gaps exceeding a consecutive hour (Long-gap events), the entire gap is filled using the climatological hourly mean computed exclusively from the training period, combined with spatiotemporal information from neighboring stations — a fixed constant that carries no information from any test-period observation and therefore introduces no look-ahead contamination regardless of the gap's position in the series.

We inspected every short-gap interval (<1-hour) at all three stations and determined, for each, whether its right-side anchor crossed into a later calendar hour. The results are summarised in Table S2.

Table S2. Short-gap interpolation events and cross-hour anchor incidence in the test period.

Station	Short-gap events	Long-gap events	Affected hourly samples (% of test set)
M9393	25	3	1.18
M9474	86	3	0.66
M9448	0	1	0.13

Across all three stations, the overwhelming majority of short-gap events consist of a single missing 5-minute slot. In such cases the right-side anchor is 5 minutes later and necessarily falls within the same clock hour unless the missing slot occupies the final 5-minute position of an hour (i.e. the XX:55 slot), in which case the anchor falls at XX+1:00. At sites M9393 and M9474, only 3 missing periods exceeded 1 hour, while at site M9448 there was only 1 missing period exceeding 1 hour. These missing values were filled using the intraday climatic mean calculated solely from the training period combined with spatiotemporal information from neighboring stations to cover the entire gap.

Current practice for filling short gaps in meteorological and environmental time series commonly relies on linear interpolation between the nearest valid observations before and after the gap—rather than purely backward-looking methods—because near-surface meteorological variables, including pavement temperature, vary smoothly and slowly at sub-hourly timescales. This approach is widely adopted because it generally reproduces the underlying smooth trajectory of the variable across short data gaps, making a linear blend between adjacent valid observations a physically appropriate (Karsisto, 2024; Yin et al., 2019; Darghiasi et al., 2023; Crevier and Delage, 2001; Karsisto et al., 2026). RST itself exhibits strong sub-

hourly autocorrelation, as documented in Sect. 3.1.1 (Figs. 5–6), so the information content carried by an anchor a few minutes into a later hour is, in physical terms, very close to the information already available from the last valid observation before the gap.

(3) Exclusion of imputed values from evaluation targets.

Finally, we confirm that the imputation procedures for the training and test periods are performed entirely independently. Short-gap linear interpolation within each period uses only observations from that same period; the long-gap climatological fill uses a constant computed exclusively from the training period and applied identically regardless of which period a given gap falls in. No test-period observation is used to impute any training-period value, and no information from the training period's actual realised weather is used beyond the fixed climatological constant. There is therefore no leakage between the training and test periods at any stage of the imputation pipeline. All reported evaluation metrics (MAE, RMSE, sMAPE, R^2) were computed exclusively against directly observed RST values. This ensures that the reported accuracy figures reflect model performance against genuine sensor measurements only.

Manuscript revisions

In response to this comment, we have made the following revisions:

First, The phrase "leakage-free stacking ensemble" has been removed from the Abstract and Introduction. The out-of-fold construction is now described as preventing data leakage specifically at the meta-learner training stage, which is its precise scope.

Second, Section 3.1.1 has been revised to clarify the order of operations (quality control →imputation→chronological train–test split → hourly aggregation → fold construction), to explain the hourly-aggregation argument in point (2) above.

References in support of this response:

Crevier, L. P., Delage, Y.: METRo: A new model for road-condition forecasting in Canada, J. Appl. Meteorol., 40(11), 2026 – 2037, https://doi.org/10.1175/1520-0450(2001)040<2026:MANMFR>2.0.CO;2, 2001.

Darghiasi, P., Baral, A., Mattingly, S., et al.: Estimation of Road Surface Temperature Using NOAA Gridded Forecast Weather Data for Snowplow Operations Management, J. Cold Reg. Eng., 37(4), 04023018, https://doi.org/10.1061/jcrgei.creng-691, 2023.

Karsisto, V. E.: RoadSurf 1.1: open-source road weather model library, Geosci. Model Dev., 17(12), 4837-4853, https://doi.org/10.5194/gmd-2023-175, 2024.

Karsisto, V., Hieta, L., Kämäräinen, M.: XGBoost-Based Corrections for Road Surface Temperature Forecasts, J. Appl. Meteorol. Climatol., 65(6), 881-897, https://doi.org/10.1175/JAMC-D-25-0189.1, 2026.

Yin, Z., Hadzimustafic, J., Kann, A., et al.: On statistical nowcasting of road surface temperature, Meteorol. Appl., 26(1), 1-13, https://doi.org/10.1002/met.1729, 2019.

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."

Response:

We thank the Reviewer for this technically precise comment. Following the Reviewer's request, we conducted a more rigorous quantitative comparison between the original 5-minute RST series and the hourly-resampled series used for model training. We address this in three parts: (1) a comparison of within-day variability between the two resolutions, (2) a periodogram-based significance test of dominant periodicities at both resolutions, and (3) an explanation of why neither finding affects the reported model comparisons, given that all evaluated models use the identical hourly dataset.

(1) Comparison of 24-hour variance between 5-minute data and smoothed 1-hour data.

To assess whether hourly averaging measurably alters the magnitude of genuine RST variability, we computed the within-day variance of RST separately for the 5-minute series (288 valid samples per day) and the hourly series (24 valid samples per day), restricted to the 361 days for which both series had complete, unimputed coverage within the same calendar day. This day-level unit places the two resolutions on an identical statistical basis, in contrast to comparing a single absolute value to a multi-hour aggregate.

The mean daily variance is 16.650°C² for the 5-minute series and 17.164°C² for the hourly series, a ratio of 1.031. This near-unity ratio indicates that hourly averaging preserves the day-to-day magnitude of genuine RST variability rather than systematically suppressing or inflating it. Fig. S1 presents the two resulting daily variance series plotted across the full study period for M9393 site, allowing direct visual comparison of their temporal evolution; the two series track each other closely throughout, with no systematic divergence at any point in the four winter seasons. Similar conclusions hold at the other two stations: at M9474, the mean daily variance is 6.856°C² for the 5-minute series and 7.042°C² for the hourly series (ratio 1.027), and at M9448, the corresponding values are 19.380°C² and 19.925°C² (ratio 1.028). Across all three stations, the hourly-to-5-minute variance ratio falls within a narrow band of 1.027–1.031, consistent with a small and station-invariant smoothing effect that does not alter the overall magnitude or temporal pattern of RST variability.

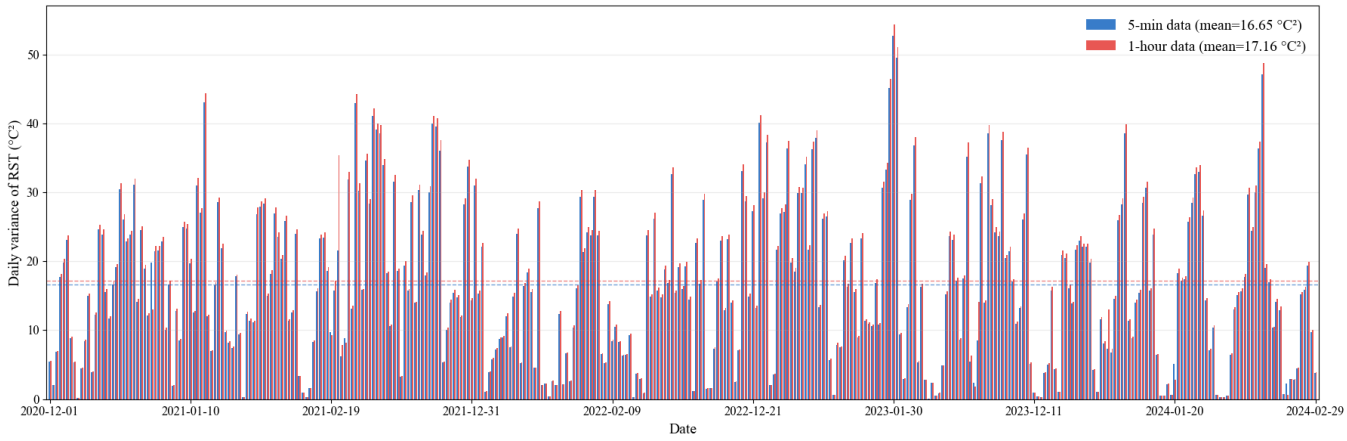

Figure S1. Bar chart of daily variance for M9393 site 5-minute raw data and 1-hour data.

Table S3. The mean daily variance values and their ratios across all three stations.

Station	5-min mean daily variance (° C²)	1-hour mean daily variance (° C²)	N days
M9393	16.650	17.164	361
M9474	6.856	7.042	361
M9448	19.380	19.925	330

(2) Aliasing artifacts: periodogram analysis with red-noise significance testing.

To directly test for aliasing-induced spurious periodicity, we computed the periodogram of the longest continuous winter segment of both the 5-minute series (26,208 samples; 2,184 hours) and the corresponding hourly series (2,184 samples), and

assessed the statistical significance of spectral peaks against a first-order autoregressive (AR(1), "red noise") background spectrum, following the standard periodogram significance-testing procedure used in climate time series analysis (Wei, 2007). For each series, the AR(1) coefficient was estimated from the lag-1 autocorrelation, the corresponding theoretical red-noise spectrum was constructed, and an F-distribution-based 95% confidence threshold was applied to identify statistically significant periods.

The two series exhibit closely matching significant periodicities. For the 5-minute series, the dominant significant periods are 24.00 h, 28.00 h, 21.00 h, and 12.00 h; for the hourly series, the dominant significant periods are identical: 24.00 h, 28.00 h, 21.00 h, and 12.00 h. The 24-hour diurnal cycle is, as expected, the single most prominent significant peak in both series, with the 12-hour semi-diurnal harmonic also clearly significant in both, consistent with the two-phase (daytime heating, nocturnal cooling) structure of RST evolution documented in Sect. 3.1.1. The AR(1) coefficients are similarly high for both series (5-minute: $r = 0.9997$; 1-hour: $r = 0.979$), reflecting the strong autocorrelation of RST at both timescales. This correspondence between the significant periodicities of the two series is reproduced at the other two stations. At M9474, the dominant significant periods for the 5-minute series are 24.00 h, 28.00 h, 21.00 h, 84.00 h, and 12.00 h; for the hourly series, the four leading periods are 24.00 h, 28.00 h, 21.00 h, and 12.00 h, with the 84-hour period absent from the hourly series owing to its position near the low-frequency boundary of the periodogram rather than any aliasing effect. At M9448, the significant periodicities are 24.00 h, 28.00 h, 21.00 h, 12.00 h, and 12.92 h for both resolutions, identical to M9393. Across all three stations, the 24-hour diurnal and 12-hour semi-diurnal peaks appear as the consistently dominant periodicities in both the 5-minute and hourly series, and no spurious low-frequency periodicity is introduced by hourly averaging that is absent from the native-resolution series. This cross-station consistency confirms that hourly block-averaging preserves the physically meaningful periodic structure of RST without introducing aliasing-induced artefacts.

Critically, no period present as significant in the hourly series is absent from, or shifted relative to, the corresponding significant period in the 5-minute series, and no spurious low-frequency period emerges in the hourly series that is not already present in the 5-minute series. This is the diagnostic signature expected in the absence of aliasing: downsampling via hourly block-averaging—an inherent low-pass (anti-aliasing) operation—preserves the genuine periodic structure of RST rather than folding unresolved high-frequency content into spurious low-frequency peaks. Fig. S2 presents the periodograms with the AR(1) significance threshold for both resolutions; the close correspondence between panels (a) and (b), particularly the precise alignment of the 24-hour and 12-hour peaks, provides direct visual confirmation of this conclusion.

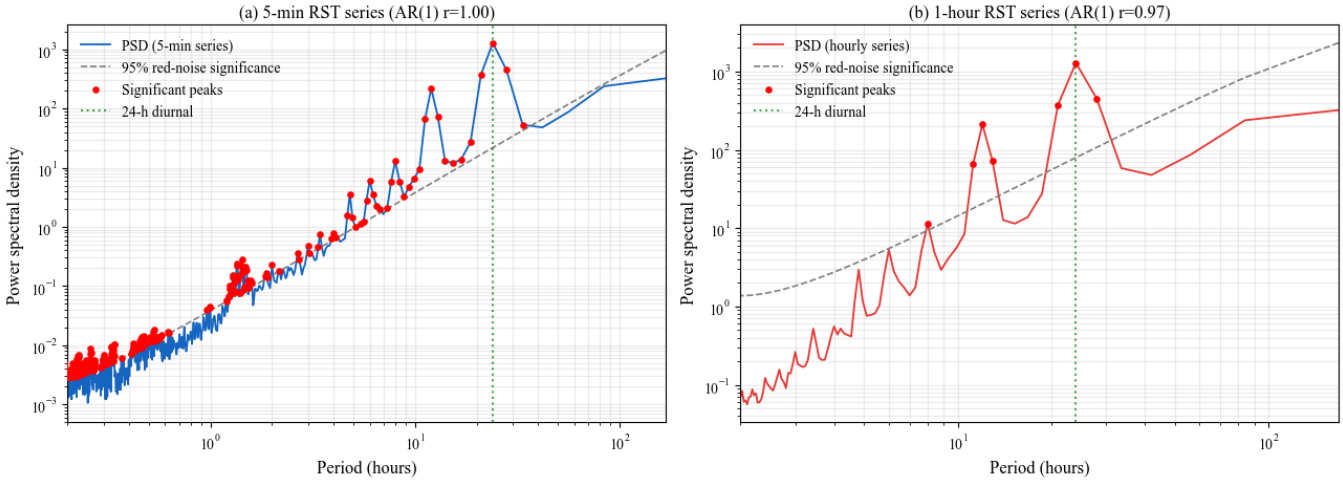

Figure S2. Power spectral density plots of 5-minute raw data and 1-hour data at M9393 site.

(3) Why this concern does not affect the reported model comparisons?

We further note that, irrespective of the magnitude of smoothing or the absence of aliasing artifacts established above, this concern does not affect the reported model comparisons, for the following reasons.

Resolution choice as a dataset construction decision. The hourly dataset used throughout this study was constructed once, from the original 5-minute observations, as a single fixed input representation for all subsequent modeling work. We note that the choice of temporal resolution at which to construct this dataset—5, 10, 30, or 60 minutes—is in principle a free design decision; the original 5-minute characteristics of the data are not disregarded as a matter of necessity, but rather a

single hourly dataset was deliberately constructed to serve the modeling and evaluation needs of this study, and all eleven models are evaluated on this one dataset under an identical protocol. Moreover, at 5-minute resolution, the full four-winter dataset yields 102,431 raw records; constructing the LSTM input sequences with a 24-hour sliding window at this resolution would increase the number of training samples and the sequence length by a factor of 12 relative to the hourly formulation, substantially raising memory requirements and training time without a commensurate improvement in the predictive information available to the model, given that sub-hourly RST variability primarily reflects sensor-level noise rather than meteorologically driven signal.

Identical dataset across all eleven models. All eleven benchmark models evaluated in Tables 4 and 5 are trained and tested on this same hourly dataset. No model in this study has access to the 5-minute series; all models are fit to the same hourly-resampled data under the same protocol. Consequently, any resampling artifact, were one to exist, would affect all eleven models identically and cannot, by itself, generate or inflate the differential ranking among them. This argument is analogous to the standard justification for within-dataset comparisons: what matters for valid model comparison is not the absolute noise level of the data, but the consistency of the data across all compared models.

Where the core evidence of methodological contribution lies. We further clarify the respective roles of the comparisons reported in Table 5. The architectures proposed in this study are KNN-LSTM, BiLSTM-MHA, and their stacked combination ILES (the last three rows of Table 5). The primary evidence for the contribution of this work is the performance of these three models relative to the eight established reference methods (Persistence, NR, RF, XGBoost, GRU, LSTM, BiLSTM, CNN-LSTM), which represent the persistence baseline, classical regression, traditional machine learning, and standard recurrent/convolutional deep learning families. The improvement of ILES over its own constituent base learners (KNN-LSTM and BiLSTM-MHA)—0.016°C and 0.02°C in Table 5—is a secondary, internal comparison intended to demonstrate that the stacking step adds a small additional benefit beyond either base learner alone; it is not, and was not intended to be, the primary claim of methodological superiority, and we have revised the manuscript so that this distinction is stated explicitly.

Magnitude of improvement relative to established baselines, reported as ranges. At the 1-hour horizon, ILES reduces MAE by 8.13%–32.67% relative to the six established machine-learning and deep-learning baselines (RF, XGBoost, GRU, LSTM, BiLSTM, CNN-LSTM), and by 52.96%–58.42% relative to the persistence and regression references (Persistence, NR). At the 3-hour horizon, the corresponding reduction relative to the established ML and DL baselines is 5.72%–12.73%. At the 6-hour horizon, ILES achieves lower MAE than five of the six established baselines, by margins of −5.03%–10.90%, and is marginally higher (by 5.03%) than the sixth (GRU); as with sMAPE (Sect. 4.1, revised per the Reviewer's Minor Comment 4), ILES is therefore not uniformly best on every metric and baseline at every horizon, though it remains the most consistently competitive model overall. Relative to its own constituent base learners, ILES further reduces MAE by 2.86%–5.09% across the three horizons, indicating a small but directionally consistent additional gain from the stacking step.

Table S4. MAE improvement ranges of ILES against machine learning and deep learning baselines (RF, XGBoost, GRU, LSTM, BiLSTM, CNN-LSTM) across forecasting horizons.

Horizon	Best baseline (MAE)	Worst baseline (MAE)	ILES	Improvement range
1-hour	CNN-LSTM (0.406)	XGBoost (0.554)	0.373	8.13%–32.67%
3-hour	GRU (1.345)	LSTM (1.453)	1.268	5.72%–12.73%
6-hour	GRU (2.007)	LSTM (2.366)	2.108	-5.03%–10.90%

Taken together, the magnitude and consistency of the improvement over the eight established reference methods—rather than the smaller margin over the two base learners proposed in this study—constitutes the principal evidence supporting the methodological contribution of this work.

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."

Response:

We thank the Reviewer for this careful and precise reading. The Reviewer correctly identifies an inconsistency between the MIMO description added in the previous revision (Sect. 2.5, ll. 289–302) and the actual implementation, in which all base learners use a single-unit dense output layer and produce scalar predictions.

We acknowledge that the MIMO characterisation introduced in the previous revision did not accurately reflect the implemented architecture. This characterisation appears to have been introduced when responding to a prior reviewer comment on multi-horizon forecasting strategy, without verifying consistency against the actual implementation. We thank the Reviewer for identifying this inconsistency and apologise for the resulting confusion.

We have re-examined the implementation in detail and confirm the following. The framework employs the direct multi-step strategy (Taieb et al., 2012): three fully independent instances of the ILES ensemble—each comprising its own KNN-LSTM, BiLSTM-MHA, and BRR meta-learner—are trained separately, one for each target horizon $\tau \in \{1, 3, 6\}$ hours. For a given horizon τ, the sliding window construction produces a scalar prediction target $y_{t+\tau}$; both base learners use a single-unit dense output layer; all weights—including the LSTM layer parameters in Eqs. 5–7 (KNN-LSTM) and the BiLSTM-MHA parameters in Eqs. 9–13—are independently optimised for that specific horizon, with no parameter sharing across the three horizon-specific instances. The BRR meta-learner for horizon τ receives a two-dimensional out-of-fold prediction vector (one scalar prediction per base learner, both for horizon τ) and produces a scalar ensemble output for that horizon. Three fully independent ILES instances are therefore reported in Sect. 4.1 (Table 5), one per forecasting horizon.

Manuscript revisions

We have made the following revisions to ensure full consistency between the methodology description, the architecture specification, and the actual implementation:

1. Section 2.5 has been rewritten to correctly describe the direct strategy as the multi-horizon forecasting approach adopted in this study, including a brief comparison with the recursive and MIMO alternatives, and the rationale for adopting the direct approach—namely, that horizon-specific optimisation avoids the output-layer distribution mismatch that can arise when a single model is required to simultaneously minimise losses at horizons with substantially different error magnitudes (Taieb et al., 2012).
2. Section 2.1 (following Eq. 8) and Section 2.2 (following Eq. 13) now each include an explicit clarifying sentence: "A separate instance of this model, with independently optimised weights $\theta_1, \theta_2, \theta_3$ (for KNN-LSTM) / all BiLSTM-MHA parameters (for BiLSTM-MHA), is trained for each forecasting horizon; no parameters are shared across horizons."
3. Section 2.4 now states: "A separate BRR meta-learner is fitted independently for each forecasting horizon, receiving the two-dimensional out-of-fold prediction vector generated by the two base-learner instances trained for that horizon."
4. Section 3.3.2 (training hyperparameters) has been revised to clarify that the architectural configuration (layer counts, hidden units, attention heads, optimizer settings) is held fixed across the three forecasting horizons, while each horizon-specific instance is trained independently with its own weights and no parameter sharing, consistent with the direct strategy described in Sect. 2.5.

These revisions ensure full consistency among the methodology description (Sects. 2.3–2.5), the model architecture specifications (Sects. 2.1–2.2), and the training protocol (Sect. 3.3.2), and accurately reflect the implementation used to generate all reported results. We confirm that all three horizon-specific model instances were trained completely independently, with no shared weights at any stage. No model was retrained in response to this comment; the architecture description has been corrected to match the implementation that produced the results already reported, and all reported results (Tables 4–7, Figs. 8–15) are therefore unchanged.

References of this response:

Taieb, S. B., Bontempi, G., Atiya, A. F., et al.: A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition, Expert Syst. Appl., 39(8), 7067-7083, https://doi.org/10.1016/j.eswa.2012.01.039, 2012.

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."*

Response:

We thank the Reviewer for these precise and important observations. We agree that the proxy argument was not rigorous enough and that the SHAP analysis required clarification. We have revised the manuscript in three respects.

(1) Tempering the proxy argument.

We have revised all passages that implied historical RST and air temperature serve as adequate proxies for incoming shortwave radiation. The revised text now consistently describes historical RST as encoding "the integrated thermal response of the pavement surface to prior meteorological forcing at the specific sensor location", without claiming that it quantifies incoming shortwave radiation, sky-view factor, local shading, or road orientation. We specifically acknowledge that similar air temperature values can correspond to substantially different RST values under different radiation and shading environments, and that the learned diurnal pattern in the input sequence reflects station-specific conditions rather than a general radiation-driven forcing signal. The following passages have been revised accordingly: the RST characteristic description (lines 329–330), the feature selection rationale (lines 383–384), the sliding window motivation (lines 404–407), and the error structure discussion (lines 565–567).

(2) Station-specific scope.

We have added explicit language in both the multi-site validation section and conclusion section to clarify the station-specific scope of the framework.

In Sect. 4.4 (multi-site validation), the following sentence has been added to the opening paragraph: "As with the primary station M9393, both M9474 and M9448 are situated in relatively unobstructed settings without significant roadside screening structures; this validation therefore assesses transferability of the framework's architecture and training protocol across stations sharing broadly similar exposure geometries, rather than across the specific shading or sky-view conditions encountered at screened road segments."

In the Conclusion (Sect. 5), the limitations paragraph has been revised from the previous text to read: "The framework was evaluated under a temperate monsoon climate, and its transferability to subarctic, alpine, or maritime regimes remains to be assessed. The absence of direct shortwave radiation measurements and local exposure descriptors—such as sky-view factor, roadside screening geometry, and road orientation—means that the model is best characterised as a station-specific time-series

predictor at a fixed exposure. The learned diurnal pattern in the RST input sequence is conditioned on the specific radiation environment at each monitored cross-section and should not be extrapolated to nearby road segments with substantially different shading or orientation without site-specific retraining or the inclusion of explicit radiation inputs."

(3) SHAP analysis scope clarification.

The Reviewer correctly observes that historical RST is included as a model input but does not appear in the SHAP plots (Figures 13 and 14), which display only the four instantaneous meteorological variables: air temperature (AT), relative humidity (RH), wind speed (WS), and precipitation (P).

We clarify that this is a deliberate analytical choice, not an omission. The historical RST input enters the model as a 24-step temporal sequence (one value per hour over the 24-hour input window), yielding 24 separate feature channels. In the KernelSHAP framework applied here, attributing importance to individual time steps within this sequence produces a set of 24 SHAP values whose interpretation is ambiguous—it is not meaningful to assert that "RST at t−3 hours" is more or less important than "RST at t−12 hours" without a separate temporal attribution study. Aggregating these 24 values into a single "lagged RST importance" figure would obscure the temporal structure and conflate distinct causal mechanisms. We have therefore restricted the SHAP analysis to the four instantaneous meteorological variables, for which each SHAP value has a clear and unambiguous physical interpretation. We note that a time-aggregated RST importance (summed across the 24 input time steps) was computed as a diagnostic check and was found to be the dominant contributor, as expected given RST's strong autocorrelation; however, because this aggregate conflates distinct temporal mechanisms, we do not present it as part of the primary SHAP analysis and instead note it here for completeness.

We have added the following clarification to the SHAP analysis section:

"The SHAP analysis presented here is restricted to the four instantaneous meteorological input variables (AT, RH, WS, P). The reported SHAP importances therefore characterise the marginal contributions of the contemporaneous meteorological state, not the full input space."

We further clarify that the dominance of AT in the SHAP analysis demonstrates a strong statistical association between air temperature and RST at this station under this exposure geometry, consistent with surface energy balance theory. It does not imply that solar radiation forcing is adequately represented; the absence of direct radiation measurement is acknowledged as a limitation.

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."*

Response:

We thank the Reviewer for this important conceptual clarification. We agree that the scope of the uncertainty quantification was not precisely stated, and that the distinction between meta-learner-level posterior predictive uncertainty and physical process uncertainty is consequential given the surface energy balance framing introduced in Section 1.

The BRR meta-learner operates on the deterministic point predictions of KNN-LSTM and BiLSTM-MHA. Its posterior predictive distribution (Equation 20) captures two sources of uncertainty: (i) parameter uncertainty in the linear combination weights, propagated through the posterior covariance Σ_n; and (ii) residual aleatoric noise in the stacking layer, captured by α^{-1}. This uncertainty is conditional on the base-learner outputs and characterises the statistical reliability of the ensemble combination—it does not propagate uncertainty through the nonlinear recurrent architectures of the base learners, and it does not model uncertainty in the physical processes governing RST evolution, such as unobserved shortwave radiation, pavement thermal conductivity, surface moisture state, or freezing-related phase transitions.

We acknowledge that the original manuscript did not make this distinction clearly enough, particularly given the prominent role of the surface energy balance framework in motivating the model design.

Manuscript revisions

We have made the following targeted revisions:

1. Abstract (lines 25–27): The phrase "probabilistic forecasts with closed-form uncertainty quantification" has been revised to "probabilistic forecasts with closed-form posterior predictive uncertainty at the ensemble combination layer," making the scope of the uncertainty claim explicit at first mention.
2. Section 1 (lines 125–129): The original sentence “ILES integrates two architecturally complementary base learners—KNN-LSTM for similarity-driven local pattern retrieval and BiLSTM-MHA for long-range temporal dependency extraction—within a leakage-free stacking ensemble, where a Bayesian Ridge Regression meta-learner fuses the base learner outputs to yield probabilistic forecasts with closed-form uncertainty quantification” is modified to “ILES integrates two base learners with partially complementary predictive characteristics—KNN-LSTM for similarity-driven local pattern retrieval and BiLSTM-MHA for long-range temporal dependency extraction—within a stacking ensemble, where a Bayesian Ridge Regression meta-learner fuses the base learner outputs to yield probabilistic forecasts with closed-form posterior predictive uncertainty at the ensemble combination layer.”
3. Section 4.4 (lines 709–718): The original paragraph "In operational deployment, when the lower bound of the 95% predictive interval falls below 0°C, maintenance crews may initiate precautionary pre-treatment irrespective of the point forecast, thereby incorporating quantified uncertainty directly into decision-making. The cross-site consistency of both coverage rates and calibration shape further supports the operational transferability of the probabilistic forecasting component of ILES" has been removed in its entirety, consistent with our response to Minor Comment 1 above, which raises the related concern that the manuscript does not evaluate road-maintenance decision metrics such as hit rate, false alarm rate, or cost-loss ratios. Section 4.4 now concludes with the statement that the conservative calibration pattern is a structural property of the BRR posterior predictive decomposition (Eq. 20) rather than a site-specific artefact, without extending this finding into an operational decision-support claim.

We have ensured that all uncertainty-related claims in the manuscript are now clearly scoped to the meta-learner level, and that the manuscript no longer extends this scoped uncertainty into an unevaluated operational decision-support claim.

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."

Response:

We thank the Reviewer for this substantive comment. We respond in four parts: the feasibility of reimplementing the closest prior architectures and a review of relevant prior work; a reframing of which comparison constitutes the primary evidence of methodological contribution; supplementary statistical analyses; and the scope and evidence for the internal ensemble gains addressing the Reviewer's concern about the robustness of small absolute improvements.

(1) Review and summary of prior relevant research.

We agree that a same-dataset comparison with these architectures would be the ideal benchmark. However, faithful reimplementation is not straightforward for the following reasons. Each of these studies targets a distinct application setting characterised by a different geographic location, climate zone, and temporal resolution, with different input variable sets and evaluation periods. Complete hyperparameter specifications—particularly for the attention mechanism in Bai et al. and the ensemble weighting strategy in Dai et al.—are not fully reported in the original papers. Re-tuning these hyperparameters on our dataset would effectively produce a new model variant rather than a faithful reproduction of the original architecture as applied in its intended context; any performance difference observed would therefore conflate architectural effects with dataset and tuning effects, making the comparison difficult to interpret.

We note, however, that Table 5 already includes the fundamental architectural components underlying several of these prior methods as independent baselines evaluated under a common protocol: BiLSTM (the core architecture of Bai et al., 2022), GRU and LSTM (the components of Dai et al.'s ensemble, 2023), CNN-LSTM (Tabrizi et al., 2021), and Random Forest (the feature-selection component of Zhang et al., 2024). The comparison in Table 5 therefore evaluates the key building blocks of these prior methods in our application setting, even if it does not replicate their specific integration strategies. We have added a note to Table 1's caption clarifying that the reported accuracies across studies reflect different datasets and protocols and are presented for literature-positioning purposes only, not as directly comparable performance benchmarks.

Table 1. Summary of LSTM-based approaches for road surface temperature prediction. The accuracy values reported are drawn from the original publications and reflect different datasets, input variables, temporal resolutions, and evaluation protocols, they are presented here for literature-positioning purposes only. Abbreviations: AT, air temperature; SR, solar radiation; RH, relative humidity; P, precipitation; WS, wind speed; WD, wind direction; AP, atmospheric pressure; Depth, measurement depth below pavement surface; Time, time-of-day index.

Reference	Model	Feature	Interval	Characteristic	Evaluation
Tabrizi et al. (2021)	CNN-LSTM	RST, AT, SR	1h	CNN-LSTM hybrid architecture for multi-horizon forecasting	MAE=1.05–3.43 °C and R^2=0.98–0.80 across 1- to 6-hour horizons
Milad et al. (2021a)	Bi-LSTM	AT, Depth, Time	1h	Bidirectional LSTM with enhanced feature extraction across depth and time dimensions	MAE = 1.332 °C; R^2 = 0.956
Bai et al. (2022)	Att-BiLSTM	RST, AT, P, WS, RH	1h	Attention-augmented BiLSTM with sliding window optimisation for micro-scale prediction	MAE = 0.330 °C; 93.4% of predictions within ±1 °C
Dai et al. (2023)	GRU, LSTM	RST, AT, P, WS, RH	1h	GRU-LSTM ensemble exploiting RST periodicity and meteorological lag effects	MAE of 0.345, 0.833, and 1.743 °C at 1-, 3-, and 6-hour horizons
Zhang et al. (2024)	RF-LSTM	RST, AT, AP, WS, RH, WD	10min	RF-based feature selection integrated with LSTM for short-term prediction	MAE = 0.048 °C; 99.1% within ±0.5 °C

(2) Reframing the comparison structure.

We agree that the comparison most directly relevant to assessing the methodological value of the proposed framework is not

the small margin between ILES and its own constituent base learners, but the comparison between the three models introduced in this study and the eight established methods evaluated under an identical protocol. Table 5 reports eleven models in total: the first eight rows—Persistence, NR, RF, XGBoost, GRU, LSTM, BiLSTM, and CNN-LSTM—are established baselines drawn from prior literature; the final three rows—KNN-LSTM, BiLSTM-MHA, and ILES—are the architectures proposed in this study. The comparison between these two groups constitutes the core evidence for the methodological contribution of this work, and we have revised the manuscript throughout to make this framing explicit.

Table 5. Performance evaluation across 1-, 3-, and 6-hour forecasting intervals. The last three models (KNN-LSTM, BiLSTM-MHA, ILES) represent the novel architectures proposed in this study. The preceding eight models serve as representative baselines from established model families. All models are trained and evaluated on the same dataset under the same protocol.

Model	1-hour			3-hour			6-hour		
	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)
Persistence	0.897	1.295	35.829	2.497	3.502	72.193	4.312	5.788	101.730
NR	0.793	1.628	34.895	2.162	4.392	62.240	3.294	7.023	76.720
RF	0.524	0.779	24.777	1.415	1.975	51.337	2.259	3.281	70.928
XGBoost	0.554	0.871	26.065	1.401	1.981	50.967	2.209	3.248	70.621
GRU	0.436	0.630	22.130	1.345	1.940	51.071	2.007	2.792	67.377
LSTM	0.453	0.636	23.475	1.453	2.198	54.892	2.366	3.452	73.116
BiLSTM	0.422	0.572	22.085	1.416	1.950	53.451	2.252	3.092	72.160
CNN-LSTM	0.406	0.567	21.410	1.399	2.041	53.167	2.225	2.884	74.995
KNN-LSTM	0.389	0.536	21.348	1.285	1.926	47.433	2.170	2.942	71.853
BiLSTM-MHA	0.393	0.545	20.783	1.415	1.893	55.268	2.194	2.822	71.834
ILES	0.373	0.521	20.328	1.268	1.835	47.553	2.108	2.754	71.281

To summarise this comparison, Table S5 reports the MAE, RMSE, and sMAPE improvement of ILES relative to the six established ML and DL baselines (RF, XGBoost, GRU, LSTM, BiLSTM, CNN-LSTM) across all three forecasting horizons.

Table S5. MAE, RMSE, and sMAPE improvement of ILES relative to the established machine learning and deep learning baselines (RF, XGBoost, GRU, LSTM, BiLSTM, CNN-LSTM), across forecasting horizons.

Horizon	Metric	Baseline range (best–worst)	ILES	Improvement range
1-hour	MAE	0.406 (CNN-LSTM)–0.554 (XGBoost)	0.373	8.13%–32.67%
	RMSE	0.567 (CNN-LSTM)–0.871 (XGBoost)	0.521	8.11%–40.18%
	sMAPE	21.41% (CNN-LSTM)–26.07% (XGBoost)	20.33%	5.14%–22.02%
3-hour	MAE	1.345 (GRU)–1.453 (LSTM)	1.268	5.72%–12.73%
	RMSE	1.940 (GRU)–2.198 (LSTM)	1.835	5.41%–16.51%
	sMAPE	50.97% (XGBoost)–54.89% (LSTM)	47.55%	6.71%–13.38%
6-hour	MAE	2.007 (GRU)–2.366 (LSTM)	2.108	-5.03%–10.90%
	RMSE	2.792 (GRU)–3.452 (LSTM)	2.754	1.36%–20.22%
	sMAPE	67.38% (GRU)–75.00% (CNN-LSTM)	71.28%	−5.79%–4.95%

At the 1-hour horizon, ILES achieves an MAE of 0.37°C, compared to a range of 0.41°C (best established ML/DL baseline, CNN-LSTM) to 0.55°C (worst, XGBoost)—an improvement of 8.13% to 32.67%. RMSE ranges from 0.57°C to 0.87°C across these baselines, compared to 0.52°C for ILES (8.11%–40.18% improvement), and sMAPE ranges from 21.41% to 26.07%, compared to 20.33% for ILES (5.14%–22.02% improvement). At the 3-hour horizon, the corresponding improvement ranges are 5.72%–12.73% (MAE), 5.41%–16.51% (RMSE), and 6.71%–13.38% (sMAPE). At the 6-hour horizon, ILES achieves an MAE of 2.11°C and an RMSE of 2.75°C, within or below the established ML/DL baseline range for five of the six comparisons (improvements of up to 10.90% in MAE and 20.22% in RMSE), with the sole exception that GRU achieves a marginally lower MAE (2.01°C, i.e., ILES is 5.03% higher) and a marginally lower sMAPE (67.38% vs. 71.28%, i.e., ILES is 5.79% higher) at this horizon. We further note that ILES retains a substantial advantage over the naive

persistence baseline and the nonlinear regression reference throughout (MAE reductions of 52.96%–58.42% at 1-hour, 48.74% at 3-hour, and 51.00% at 6-hour relative to persistence), confirming that the architectural complexity of the proposed framework is well justified relative to simple reference models, even where its margin over the strongest established ML and DL baseline is comparatively modest.

We consider this comparison—spanning a consistent and, in most cases, substantial improvement across nearly all established baselines, metrics, and horizons, with one identified exception at the 6-hour horizon—to constitute the primary evidence supporting the methodological value of the proposed framework. By contrast, the 0.016–0.02°C margin between ILES and its own base learners (KNN-LSTM, BiLSTM-MHA) discussed by the Reviewer is a secondary, internal comparison, included to support the narrower claim that the stacking step provides an additional, directionally consistent gain, and should not be read as the primary basis for the paper's claim of methodological contribution. We have revised the manuscript throughout to reflect this distinction explicitly.

(3) Additional analyses addressing methodological rigour.

To address the Reviewer's concern that small absolute improvements warrant additional statistical support, we conducted supplementary significance testing on the 1-hour forecasting horizonat the M9393 site, where the margin between ILES and its base learners is smallest and therefore most in need of validation.

Bootstrap confidence intervals (10,000 resamples with the percentile method; Efron, 1979) were constructed for the test-set MAE of each model across n = 712 test predictions at the 1-hour horizon. The resulting estimates are: ILES, 0.373°C [95% CI: 0.358–0.388°C]; KNN-LSTM, 0.389°C [95% CI: 0.374–0.405°C]; BiLSTM-MHA, 0.393°C [95% CI: 0.378–0.409°C]. The confidence intervals for ILES and KNN-LSTM share a narrow overlap region (0.374–0.388°C), while those for ILES and BiLSTM-MHA do not overlap, indicating stronger statistical separation for the latter comparison.

Table S6. Paired t-tests on the absolute error series across all test samples yield the following results.

Comparison	Mean difference	t-statistic	p-value	Cohen's d
ILES vs KNN-LSTM	−0.0163°C	t = -4.899	< 0.001	-0.105
ILES vs BiLSTM-MHA	−0.0198°C	t = -7.702	< 0.001	-0.166

Both comparisons are statistically significant at $p < 0.001$. The effect sizes are modest (Cohen's d ≈ -0.1 to -0.17), which is consistent with the expected behaviour of ensemble refinements where base models have already captured much of the predictable signal and further gains reflect diminishing returns rather than a structural failure of the ensemble (Wolpert, 1992; Zhou, 2012). Small effect sizes of this magnitude are commonly observed in ensemble stacking studies and do not imply that the improvement is practically negligible; at operational timescales, a systematic reduction of 0.016– 0.020°C in the mean absolute error of winter RST forecasts represents a consistent directional gain across the full test population.

We acknowledge that this significance testing was conducted as post-hoc supplementary analysis in response to the Reviewer's comment rather than as a pre-specified component of the manuscript due to space limitations. We have therefore presented it here in the response letter rather than adding a formal significance-testing section to the main text, consistent with the manuscript's focus on the methodological framework rather than on statistical inference per se. We note, however, that the $p < 0.001$ results and the directional consistency of the ensemble gain across all three forecasting horizons (Table 5) and across all three validation stations (Table 7) collectively provide statistical and cross-site evidence that the improvement of ILES over its base learners is unlikely to be attributable to random variation in a single training run. Formal significance testing with repeated independent runs remains a direction for future work.

(4) On the internal ensemble gains: scope, evidence, and contribution.

We acknowledge that the incremental MAE gains of ILES over its strongest constituent base learner are modest in absolute magnitude. Without repeated independent training runs and formal significance testing, their statistical robustness cannot be formally established, and we do not present these margins as the primary evidence of methodological contribution.

We nonetheless regard the stacking step as a meaningful and principled component of the proposed framework, for the following reasons. First, this work makes a domain-specific methodological contribution by demonstrating that architecturally complementary LSTM-based learners can be productively combined within a leakage-free stacking framework for winter RST prediction, with the BRR meta-learner additionally providing calibrated uncertainty quantification at the ensemble combination layer—a capability not available from any of the individual baselines. Even where the absolute MAE gain is small, the

framework represents a useful integration of established components adapted to the operational requirements of this domain. Second, the directional consistency of the ensemble gain provides some structural support for its robustness: the advantage of ILES over KNN-LSTM is observed in the same direction across all three forecasting horizons in Table 5. Third, and most importantly, the framework's cross-site performance provides the strongest evidence of its practical value. When independently retrained at two additional stations with distinct geographical settings and surface conditions, ILES consistently achieves lower MAE than the standard LSTM baseline, with reductions of 17.7%, 5.7%, and 35.6% at M9393, M9474, and M9448, respectively (Table 7). This cross-site consistency indicates that the relative benefit of the proposed ensemble architecture is reproducible under site-specific retraining across the temperate monsoon climate zone examined in this study, rather than being an artefact of a single station or training run.

Manuscript revisions

We have made the following specific revisions.

First, in Table 1, we have added the following note to the caption: "The accuracy values reported are drawn from the original publications and reflect different datasets, input variables, temporal resolutions, and evaluation protocols; they are presented here for literature-positioning purposes only."

Second, in Table 5, Second, we have added an explanatory note to the Table 5 caption, distinguishing the architectures proposed in this study from the reference methods used for comparison: "The last three models (KNN-LSTM, BiLSTM-MHA, ILES) represent the novel architectures proposed in this study. The preceding eight models serve as representative baselines from established model families. All models are trained and evaluated on the same dataset under the same protocol."

Third, in Sect. 4.1, we have replaced single-point superiority claims with the range-based comparison reported in point (1) above. The revised text reports, for each horizon, the MAE reduction relative to persistence and NR, relative to the traditional machine learning and standard deep learning baselines as a group, and relative to the two base learners, expressed as ranges rather than as a single headline figure. At the 6-hour horizon, the text now explicitly notes that GRU attains a marginally lower MAE than ILES, while ILES retains the lowest RMSE among all eleven models at this horizon. Corresponding RMSE comparisons have been added in the same manner.

Fourth, throughout Sect. 4.1, we have replaced "demonstrates superiority," "robust advantage," and "consistent superiority" with measured formulations such as "achieves lower MAE and RMSE than" and "is consistent with a structural advantage of."

Finally, consistent with the revision made in response to Minor Comment 4, all claims of best performance now specify "in terms of MAE and RMSE" rather than "across all metrics," reflecting that ILES is not uniformly best in sMAPE at every horizon.

References in support of this analysis:

Bai, S., Yang, W., Zhang, M., et al.: Attention-based BiLSTM model for pavement temperature prediction of asphalt pavement in winter, Atmosphere, 13(9), 1524, https://doi.org/10.3390/atmos13091524, 2022.

Dai, B., Yang, W., Ji, X., et al.: An ensemble deep learning model for short-term road surface temperature prediction, J. Transp. Eng. Part B-Pavements, 149(1), 04022067, https://doi.org/10.1061/JPEODX.PVENG-1215, 2023.

Efron, B.: Bootstrap methods: Another look at the jackknife, Ann. Stat., 7(1), 1–26, https://doi.org/10.1214/aos/1176344552, 1979.

Milad, A., Adwan, I., Majeed, S. A., et al.: Emerging technologies of deep learning models development for pavement temperature prediction, IEEE Access, 9, 23840-23849, https://doi.org/10.1109/ACCESS.2021.3056746, 2021a.

Tabrizi, S. E., Xiao, K., Thé, J. V. G., et al.: Hourly road pavement surface temperature forecasting using deep learning models, J. Hydrol., 603, 126877, https://doi.org/10.1016/j.jhydrol.2021.126877, 2021.

Wolpert, D. H.: Stacked generalization, Neural Netw., 5(2), 241-259, https://doi.org/10.1016/S0893-6080(05)80023-1, 1992.

Zhang, M., Guo, H., Li, J., et al.: A deep learning approach for enhanced real-time prediction of winter road surface temperatures in high-altitude mountain areas, Promet-Traffic&Transportation, 36(5), 958-972, https://doi.org/10.7307/ptt.v36i5.525, 2024.

Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms, CRC Press, Boca Raton, https://doi.org/10.1201/b12207, 2012.

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."*

Response:
We thank the reviewer for this precise observation. We agree that the manuscript does not evaluate road-maintenance decision metrics such as hit rate, false alarm rate, miss rate, or cost-loss ratios. Rather than retaining the operational claim with additional qualification, we have removed the relevant passage from Sect. 4.4 in its entirety, retaining only the description of the reliability diagrams, the empirical coverage statistics, and the cross-site consistency of the calibration pattern itself.

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."*

Response:
We thank the Reviewer for noting these issues. We have proofread the entire manuscript line by line and corrected all identified errors, including "focasting" → "forecasting" (two instances, Sect. 4.1), and have standardised the wording and formatting of all figure and table captions for internal consistency (e.g. consistent capitalisation of variable names, and consistent phrasing of "forecasting horizon" throughout). We have additionally performed a full pass for awkward phrasing flagged by the Reviewer's general comment and corrected instances identified during this review.

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."*

Response:
We thank the Reviewer for this important clarification. The studies summarised in Table 1 were conducted on different datasets, with different input variables, temporal resolutions, station characteristics, and evaluation protocols; direct numerical comparison of the reported accuracy values across studies is therefore not meaningful, and Table 1 should be read as a literature-positioning summary rather than a benchmark comparison. We have added the following note to the Table 1 caption: "The accuracy values reported are drawn from the original publications and reflect different datasets, input variables, temporal resolutions, and evaluation protocols, they are presented here for literature-positioning purposes only."

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."*

Response:

The Reviewer correctly identifies several instances where the manuscript overstates the uniformity of the performance advantage of ILES and Variable combination 3. We confirm the following discrepancies: at the 3-hour horizon, KNN-LSTM achieves a lower sMAPE than ILES (47.433% vs. 47.553%, Table 5); at the 6-hour horizon, GRU, XGBoost, and RF all achieve lower sMAPE than ILES (67.377%, 70.621%, and 70.928% vs. 71.281%, Table 5); Variable combination 3's 6-hour sMAPE (100.355%, Table 6) is higher than those of combinations 1 (91.885%) and 2 (97.241%); and ILES does not achieve the best sMAPE at station M9474 (Table 7). We have revised every instance of "across all metrics and horizons" (and equivalent absolute phrasing) to accurately distinguish MAE/RMSE rankings from sMAPE rankings. The specific revisions are listed below, organised by location.

Abstract. "Results indicate that ILES achieves lower prediction errors in general than ten other models" is retained as already appropriately hedged ("in general"); no change required.

Section 4.1, summary paragraph (lines 538–540). The original sentence "In summary, ILES demonstrates the most consistent and balanced performance profile across all horizons and metrics" has been revised to: " Across all three horizons, ILES achieves the lowest MAE and RMSE among the eleven evaluated models, with the single exception of GRU's marginally lower MAE at 6 hours. sMAPE rankings are less consistent: KNN-LSTM attains a marginally lower sMAPE at 3 hours, and GRU, XGBoost, and RF attain lower sMAPE at 6 hours."

Section 4.2, opening of the input-configuration comparison (lines 596–597). The original sentence "As shown in Table 6, Variable combination 3 achieves the lowest errors across all metrics and forecasting horizons" has been revised to: "As shown in Table 6, Variable combination 3 achieves the lowest MAE and RMSE across all three forecasting horizons. Its sMAPE is also lowest at the 1-hour and 3-hour horizons."

Conclusion (Section 5) (lines 723–746). The original sentence "ILES achieved the lowest errors across all forecasting horizons and evaluation metrics" has been revised to: " ILES generally achieved the lowest prediction errors among all eleven evaluated models at each of the 1-, 3-, and 6-hour forecasting horizons." And the original sentence "physics-motivated feature engineering incorporating the surface–air temperature difference and multi-scale RST temporal tendency features consistently outperformed both the station-only baseline and ERA5-Land reanalysis augmentation across all forecasting horizons and meteorological regimes" has been revised to: " physics-motivated feature engineering incorporating the surface–air temperature difference and multi-scale RST temporal tendency features overall consistently outperformed both the station-only observational baseline and ERA5-Land reanalysis augmentation across all forecasting horizons and meteorological regimes."

We have additionally performed a full-text search for the phrases "across all metrics," "across all horizons," "best across," and "all metrics and horizons" to confirm that no remaining instance overstates uniformity beyond what Tables 5–7 support.

Dear Reviewer:

We thank the Reviewer for the positive overall assessment and for identifying these two remaining issues. We address each in turn.

1. ABLATION INTERPRETATION (Section 3.4, Table 4)

"The ablation study reports an MAE reduction of 0.035°C (Config-2 vs Config-1) from adding multi-head attention alone, and 0.044°C (Config-3 vs Config-1) from adding KNN-based similarity augmentation alone. The combined ILES configuration achieves an MAE reduction of 0.054°C. Since 0.054 < 0.035 + 0.044 = 0.079, the combined gain is sub-additive, indicating that the two components share part of their effective contribution rather than synergizing. The manuscript text in Section 3.4 correctly identifies this as partial overlap. The authors should ensure that: (a) The interpretation of "complementarity" in the abstract, introduction, and ensemble motivation sections is consistent with this sub-additive result. Phrases such as "architecturally complementary" should be moderated to "partially complementary" or similar. (b) No claim of "synergistic interaction" appears in the manuscript or the supplementary material. The ensemble approach is still supported by the fact that ILES outperforms all three individual configurations, but the strength of the complementarity claim should match the data."

Response:

We agree with the Reviewer's observation that the sub-additive combined gain (0.054°C < 0.035 + 0.044 = 0.079°C) indicates partial overlap in the effective contributions of the two components, and that the term "architecturally complementary" overstates the nature of this relationship.

We have conducted a systematic search of the full manuscript and revised all instances of complementarity language to be consistent with the ablation result. Specifically:

The phrase "architecturally complementary base learners" in the abstract and introduction has been revised to "base learners with partially complementary predictive characteristics." We have also added a forward reference to the ablation result in the introduction, noting that the combination yields sub-additive rather than synergistic gains. The phrase "complementary strengths" in Sections 4.1 and elsewhere has been revised to "partially non-overlapping predictive profiles." We have verified that no instance of "synergistic" or "synergy" appears in the manuscript or supplementary material. Section 3.4 already correctly identified the sub-additive nature of the combined gain; we have ensured that the concluding sentence of this section uses the formulation "partially complementary rather than redundant or synergistic contributions," consistent with the ablation data.

The ensemble approach remains well-supported: ILES outperforms all three ablation configurations, and the multi-site consistency of this advantage across three stations and three forecasting horizons provides additional evidence that the improvement is structurally stable. The revised text reflects this more accurately than the original "complementary" framing.

2. SCOPE OF GENERALIZABILITY CLAIMS

"*The multi-site validation at M9474 and M9448 strengthens the manuscript, but all three stations lie within Jiangsu Province in the same temperate monsoon climate zone. The Conclusion acknowledges this. The abstract and introduction should be adjusted accordingly: phrases such as "transferable across different road environments" should be qualified as "transferable across different road environments within the temperate monsoon climate zone."*

Response:

We agree that "transferable across different road environments" is too broad given that all three validation stations lie within Jiangsu Province in the same temperate monsoon climate zone.

We have revised the relevant passages in the abstract, introduction, Section 4.4, and conclusion as follows.

In the abstract, the phrase "confirming the transferability of the proposed framework across different road environments" has been revised to "confirming the transferability of the proposed framework across different road environments within the temperate monsoon climate zone."

In Section 4.4, the concluding sentence of the multi-site validation discussion now reads: "The consistent performance advantage of ILES across stations with distinct surface conditions confirms the structural robustness of the proposed approach within the temperate monsoon climate zone of Jiangsu Province."

In the conclusion, the corresponding claim has been similarly qualified. The existing limitation statement regarding climate transferability (previously in the final paragraph only) has been moved forward to the multi-site validation summary to ensure that the scope is stated at the point where the generalizability claim is made, not only in the limitations section.

We believe these revisions fully address both remaining concerns and are prepared to provide a clean revised manuscript for the Reviewer's consideration.