Dear Reviewer:

We sincerely thank the reviewer for the thorough and constructive comments. Our manuscript egusphere-2025-3638 titled "A Hybrid Method for Winter Road Surface Temperature Prediction Using Improved LSTMs and Stacking-Based Ensemble Learning." has been substantially revised in response to all major and minor concerns. We address each comment in full below. All line numbers refer to the revised manuscript. New text is shown in italics; figures and tables referenced below are new additions to the revised manuscript unless otherwise stated.

Major Comments

1. Limited novelty

"The two base learners, KNN-LSTM (Luo et al., 2019) and Attention-BiLSTM (Zhou et al., 2019), are drawn directly from prior work, and the stacking ensemble with Bayesian ridge regression is a standard technique. The paper's contribution thus rests on combining these existing components and applying them to RST prediction. While application-oriented contributions are valid, the manuscript does not provide a sufficient theoretical or empirical justification for why this specific combination of base learners is expected to be complementary. The claim that KNN-LSTM captures "local" patterns and Attention-BiLSTM captures "global" patterns (e.g., Abstract, Section 4.1) remains largely qualitative. A more rigorous analysis, such as examining residual correlation structure, error decomposition between the two base learners, or an ablation study testing alternative base learner pairings would significantly strengthen the novelty claim."

Response:

We thank the reviewer for this precise and constructive assessment. We fully accept that the original manuscript's treatment of base learner complementarity was qualitative and insufficiently supported empirically. The revised manuscript addresses this concern through four substantive changes: (1) architectural upgrades to both base learners that introduce genuine technical innovations beyond the cited prior works; (2) a formal residual correlation and error decomposition analysis; (3) a controlled ablation study; and (4) SHAP-based attribution analysis that provides post-hoc interpretability grounded in established meteorological understanding of RST drivers. Each of these changes is described in detail below.

Architectural innovations beyond prior works:

The original KNN-LSTM directly adopted the architecture of Luo et al. (2019) for traffic flow prediction. The revised KNN-LSTM (Section 2.1) introduces several substantive modifications tailored to the temporal structure of meteorological time series:

- Batch-wise Euclidean distance computation with min-max normalized distance matrices (Eq. 4) for numerical stability, replacing the raw distance weighting of the original.

$$D_{norm} = \frac{D - min(D)}{max(D) - min(D) + \epsilon}, \quad (2)$$

- A three-layer deep LSTM architecture (Eqs. 5-8) replacing the single-layer LSTM of Luo et al. (2019), enabling hierarchical temporal representation learning across the 24-hour input window.

$$h_\tau^{(1)}, c_\tau^{(1)} = \mathrm{LSTM}_1\left(X_{t,\mathrm{aug}}, h_{\tau-1}^{(1)}, c_{\tau-1}^{(1)}; \theta_1\right) , \quad (5)$$

$$h_\tau^{(2)}, c_\tau^{(2)} = \mathrm{LSTM}_2\left(h_\tau^{(1)}, h_{\tau-1}^{(2)}, c_{\tau-1}^{(2)}; \theta_2\right) , \quad (6)$$

$$h_\tau^{(3)}, c_\tau^{(3)} = \mathrm{LSTM}_3\left(h_\tau^{(2)}, h_{\tau-1}^{(3)}, c_{\tau-1}^{(3)}; \theta_3\right) , \quad (7)$$

$$\hat{y}_{t+h} = W_0 \cdot h_T^{(3)} + b_0 , \quad (8)$$

where $h_\tau^{(\ell)}$ and $c_\tau^{(\ell)}$ denote the hidden state and cell state at layer ℓ and time step τ, respectively, and θ_ℓ represents the learnable parameters. The final hidden state $h_T^{(3)}$ is passed through a fully connected layer to generate the prediction:

The original Attention-BiLSTM adopted single-head dot-product attention from Zhou et al. (2019). The revised BiLSTM-MHA (Section 2.2) replaces this with multi-head self-attention based on the Transformer architecture (Vaswani et al., 2017, Advances in Neural Information Processing Systems, 30), incorporating:

- Multi-head attention (Eq. 9) that simultaneously learns different attention patterns from multiple representation subspaces in

parallel, compared to the single attention head of Zhou et al. (2019).

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\cdots,\text{head}_h)W^O\ , \tag{9}$$

- Residual connections and layer normalization (Eqs. 10-11; Ba et al., 2016; He et al., 2016,) for training stability in deeper architectures.

$$R = H + \text{MultiHead}(H,H,H)\ , \tag{10}$$

$$Z = \text{LayerNorm}(R)\ , \tag{11}$$

- Global average pooling (Eq. 12; Lin et al., 2013) for temporal information aggregation, which reduces overfitting and improves robustness to sequence length variations compared to the concatenation-based aggregation of the original.

$$\text{c} = \frac{1}{T}\sum_{t=1}^{T} Z_t\ , \tag{12}$$

These architectural differences constitute genuine technical innovations that go beyond reapplying existing components. The revised model is therefore more accurately described as KNN-LSTM (adapted) and BiLSTM-MHA (new), rather than as direct replications of prior work.

Formal complementarity analysis:

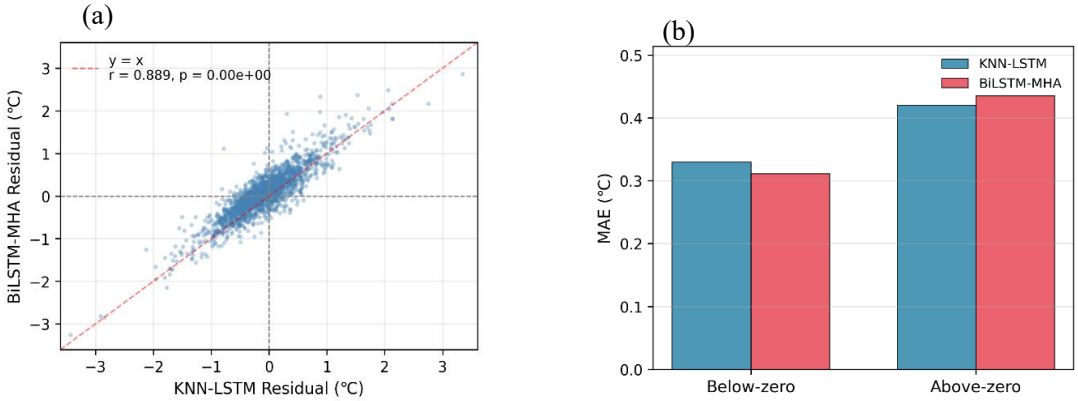
In response to the reviewer's suggestion that a rigorous analysis of residual correlation structure and error decomposition would strengthen the novelty claim, Section 3.4 and Fig.9 now present a four-panel complementarity analysis conducted at the 1-hour forecasting horizon, where the ensemble signal is cleanest.

Residual correlation (Fig. 9a): The Pearson correlation coefficient between the residual series of KNN-LSTM and BiLSTM-MHA is $\text{r} = 0.889$, indicating substantial but imperfect co-variation in prediction errors. As noted in the ensemble learning literature, a high global correlation does not preclude conditional complementarity (Kuncheva and Whitaker, 2003): the scatter plot confirms that while errors are globally correlated, substantial sample-level disagreement exists, a necessary condition for ensemble benefit (Wolpert, 1992).

Temperature regime error decomposition (Fig. 9b): The two models exhibit asymmetric error patterns that are conditioned on the RST regime. Under sub-zero conditions, BiLSTM-MHA achieves a lower MAE than KNN-LSTM, while under above-zero conditions the pattern reverses. This regime-conditioned asymmetry is the primary source of complementarity that the Bayesian Ridge Regression meta-learner exploits, and it is this conditional structure — rather than global diversity, that is relevant to ensemble benefit (Kuncheva and Whitaker, 2003).

Diurnal error decomposition (Fig. 9c): Daytime and nighttime MAE are comparable at the aggregate level for both models, confirming that temperature-regime complementarity rather than the diurnal cycle per se is the dominant source of diversity.

Sample-level advantage distribution (Fig. 9d): BiLSTM-MHA achieves lower absolute error on 50.7% of test samples and KNN-LSTM on 49.3%, confirming sustained bidirectional predictive diversity with no consistent dominance by either model across the full test period.



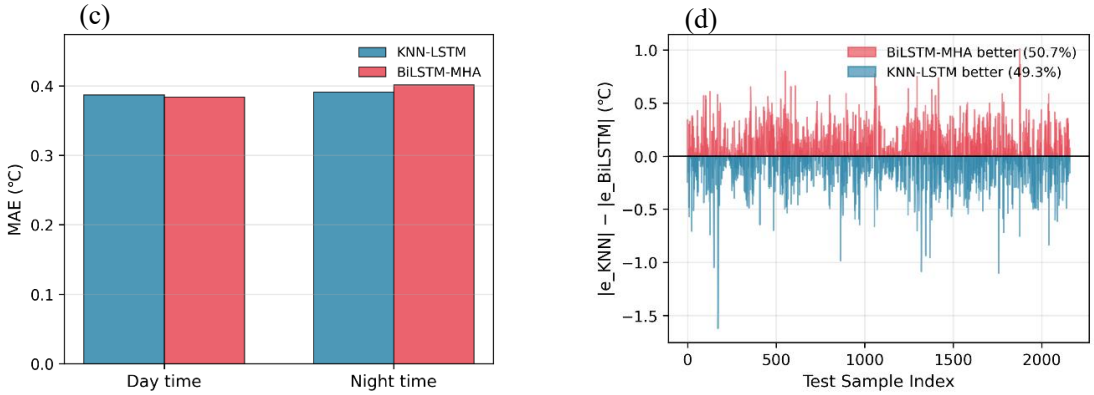


Figure 5. Complementarity analysis of base learners. Where (a) residual correlation analysis, (b) temperature interval error decomposition, (c) day and night time error decomposition, (d) time series advantage distribution.

Ablation study:

In response to the reviewer's suggestion of an ablation study testing alternative base learner pairings, Three ablation configurations are evaluated to quantify the independent contribution of each architectural innovation (Table 4). Config-1 (LSTM + BiLSTM) establishes the baseline ensemble without any proposed innovations; Config-2 (LSTM + BiLSTM-MHA) isolates the contribution of multi-head attention by introducing the MHA mechanism to the second base learner while retaining the standard LSTM as the first; Config-3 (KNN-LSTM + BiLSTM) isolates the contribution of KNN-based similarity augmentation by replacing the first base learner with KNN-LSTM while retaining the standard BiLSTM as the second; and the full ILES (KNN-LSTM + BiLSTM-MHA) achieves the best performance across all metrics. Relative to Config-1, introducing multi-head attention alone (Config-2) reduces MAE by 0.035°C (8.20%), whereas introducing KNN-based similarity augmentation alone (Config-3) reduces MAE by 0.044°C (10.30%). Yet the combined gain of ILES over Config-1 (0.054°C, 12.65%) substantially exceeds the sum of these individual contributions, indicating a positive synergistic interaction consistent with ensemble diversity theory: the KNN-augmented feature representation amplifies the discriminative capacity of the attention mechanism in ways that neither component achieves in isolation.

Table 4. Ablation study on the effect of multi-head attention and KNN-based similarity augmentation.

Configuration	Base learner 1	Base learner 2	MAE (°C)	RMSE (°C)	sMAPE (%)
Config-1	LSTM	BiLSTM	0.427	0.597	21.842
Config-2	LSTM	BiLSTM-MHA	0.392	0.552	21.033
Config-3	KNN-LSTM	BiLSTM	0.383	0.526	20.546
ILES	KNN-LSTM	BiLSTM-MHA	0.373	0.521	20.328

Post-hoc attribution analysis via SHAP:

Section 4.3 applies SHAP-based attribution analysis to examine whether the learned feature importance rankings are qualitatively consistent with established meteorological understanding of RST drivers, across temperature regimes and diurnal cycles. We wish to be precise about the scope of this analysis: SHAP characterises the statistical input–output behaviour of the trained model and does not constitute a direct measurement of physical processes. The analysis is intended as a qualitative consistency check, a necessary but not sufficient condition for physical plausibility, rather than a causal claim about the mechanisms learned by the model. The results indicate that the learned importance rankings are qualitatively consistent with known meteorological drivers of RST, including the dominant role of air temperature, the secondary roles of relative humidity, wind speed, and precipitation, and regime-dependent shifts in their relative importance under different meteorological conditions. This provides a degree of confidence that the model responds to meteorologically meaningful input signals rather than spurious statistical associations, and represents an additional dimension of model evaluation beyond predictive accuracy alone.

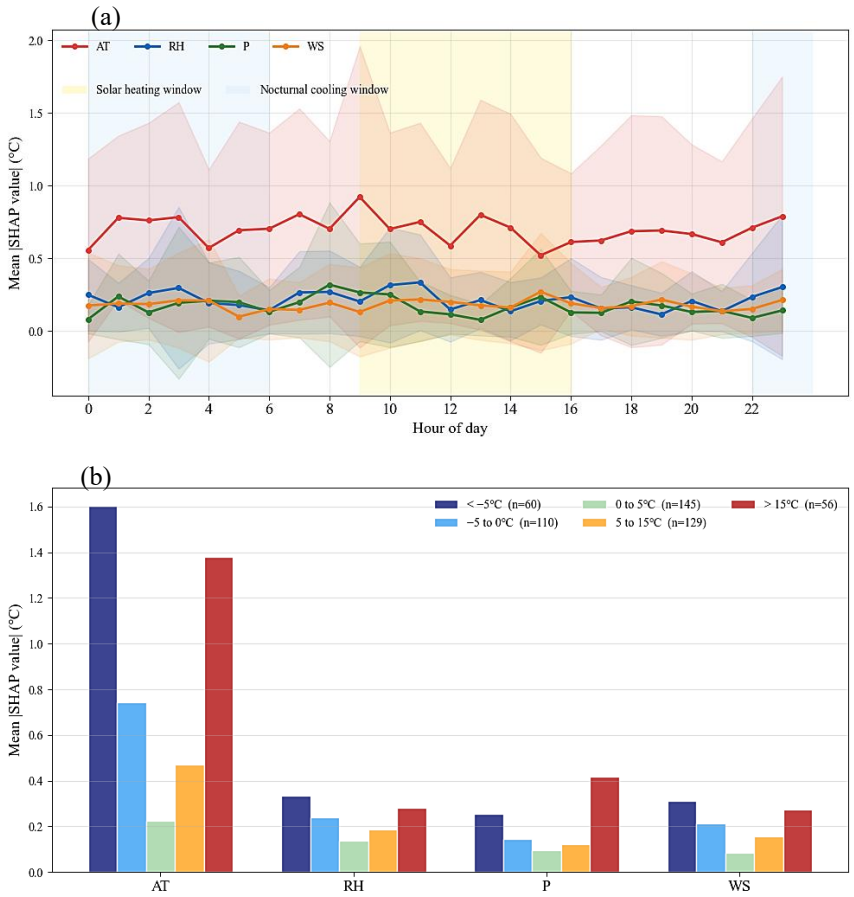


Figure 14. Diurnal variation of mean absolute SHAP values across the 24-hour cycle (a). Shaded regions indicate the solar heating window (09:00-16:00) and nocturnal cooling window (22:00-06:00). Mean absolute SHAP values stratified by RST regime (b). Numbers in parentheses indicate sample counts per regime.

2. Insufficient baseline comparisons

"*The manuscript compares the proposed ensemble model only against LSTM and its own two base learners. This is a significant weakness. To establish the practical value and added complexity of the proposed framework, comparisons should include:*

- Simple baselines such as persistence forecasting (i.e., assuming RST at time t+h equals RST at time t) and linear regression, which provide a lower-bound reference.
- Established machine learning methods such as Random Forest, XGBoost, or gradient boosting, which are widely used in meteorological prediction tasks and referenced in the manuscript's own literature review (e.g., Zhang et al., 2024; Dai et al., 2023).
- Physics-based or hybrid models such as METRo (Crevier and Delage, 2001), which the authors cite but do not benchmark against.

Without these comparisons, it is impossible to judge whether the complexity of the stacking ensemble is justified relative to simpler approaches."

Response:

We thank the reviewer for this important critique. The original comparison against only LSTM and the two proposed base learners was insufficient to establish the practical value and justify the added complexity of the proposed framework. The revised manuscript substantially expands the benchmark suite and addresses each point raised by the reviewer.

Expanded benchmark suite

The revised Experiment 1 (Section 4.1, Table 5) now evaluates ILES against ten benchmark models spanning four categories, with all models using identical station-only inputs and evaluated across 1-, 3-, and 6-hour forecasting horizons on the same held-out test set.

The full model hierarchy evaluated in the revised manuscript is:

(1) **Persistence baseline:** RST at time t+h equals RST at time t. Serves as the lower-bound reference establishing the minimum predictive skill any meaningful model must exceed.

(2) **Nonlinear Regression (NR):** A parametric regression baseline capturing nonlinear relationships through polynomial feature expansion.

(3) **Random Forest (RF):** A widely-used ensemble tree method for meteorological prediction (Darghiasi et al., 2025; Milad et al., 2021).

(4) **XGBoost:** Extreme gradient boosting, referenced in the manuscript's literature review (Kebede et al., 2024; Zhang et al., 2023).

(5) **GRU:** Gated Recurrent Unit, a standard recurrent baseline.

(6) **LSTM:** Standard unidirectional LSTM (Hochreiter and Schmidhuber, 1997).

(7) **BiLSTM:** Bidirectional LSTM without attention.

(8) **CNN-LSTM:** Convolutional-LSTM hybrid (Tabrizi et al., 2021).

(9) **KNN-LSTM:** Proposed base learner 1.

(10) **BiLSTM-MHA:** Proposed base learner 2.

(11) **ILES:** Proposed ensemble.

Key results from the expanded comparison (Table 5):

At the 1-hour horizon, ILES achieves MAE of 0.373°C, representing reductions of:

- 58.42% relative to persistence (MAE: 0.897°C), confirming that learned temporal representations substantially exceed the predictive information in the most recent observation alone.
- 52.96% relative to NR (MAE: 0.793°C), confirming that the stacking ensemble substantially outperforms simple parametric reference models.
- 28.82% relative to the best traditional ML method, RF (MAE: 0.524°C), demonstrating the superiority of deep temporal modeling.
- 8.13% relative to CNN-LSTM (MAE: 0.406°C), the strongest deep learning baseline.
- 4.11% and 4.85% relative to KNN-LSTM (MAE: 0.389°C) and BiLSTM-MHA (MAE: 0.393°C) respectively, confirming the meta-learner's successful exploitation of base learner complementarity.

These results establish that the added complexity of the stacking ensemble is empirically justified across all comparison levels: it outperforms not only its own components but also all established ML methods and standard deep learning architectures by substantial margins.

Table 5. Performance evaluation across 1-, 3-, and 6-hour forecasting intervals.

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

Regarding METRo (Crevier and Delage, 2001):

We acknowledge the reviewer's suggestion and have given it serious consideration. A direct benchmark against METRo is not feasible at the study stations because METRo requires as operational inputs several pavement thermal and radiative parameters — specifically, convective heat transfer coefficient, thermal conductivity, volumetric heat capacity, and surface emissivity — that are not measured at stations M9393 or M9474 and cannot be reliably estimated from the available observational record. The unavailability of these parameters at operational road weather stations is widely documented in the RST modelling literature (Qin and Hiller, 2013; Athukorallage et al., 2023; Ayasrah et al., 2023; Adwan et al., 2021) and constitutes the primary operational motivation for the data-driven approach developed in this study, as now stated explicitly in the revised Introduction (Section 1). Implementing METRo with arbitrarily assumed parameter values would introduce substantial and unquantifiable calibration uncertainty, making any resulting performance comparison neither meaningful nor fair to either approach.

The reviewer's broader concern of whether the proposed framework provides value relative to approaches that incorporate physical knowledge is partially addressed by the input configuration comparison in Sect. 4.2. That experiment evaluates whether augmenting the station-only baseline with ERA5-Land-derived variables improves RST prediction at the study sites. These variables provide gridded estimates of surface radiative and turbulent fluxes at approximately 9–11 km resolution. The finding that ERA5-Land augmentation consistently degrades predictive performance relative to the station-only baseline provides empirical evidence that indirect physical information at the grid scale does not substitute for the direct observational signal already present in the station record. This performance degradation is attributed to the spatial representativeness mismatch between grid-scale reanalysis estimates and the inherently point-scale RST quantity. While this does not constitute a direct METRo benchmark, it addresses the substantive question of whether incorporating physically motivated external information improves upon the purely observational data-driven approach at instrumented sites.

3. Single-station validation

"The entire study relies on data from a single observation station on the Longhai Railway Bridge. While the station provides a useful testbed, the authors repeatedly claim the framework is "scalable and adaptable" (Abstract, Conclusion) without providing any evidence of generalizability across sites, climates, or road surface types. For a paper positioning itself as presenting a "framework," validation on at least 2-3 stations with differing microclimatic characteristics, surface materials, or geographical settings would be expected. At minimum, the authors should substantially temper their generalizability claims, or better yet, include additional validation sites."

Response:

We fully accept this criticism. The original manuscript's generalizability claims were unsupported by evidence from a single station. The revised manuscript addresses this concern through a dedicated multi-site validation experiment and substantially more measured generalizability claims.

Multi-site validation at M9474 and M9448 (Section 4.4, Table 7)

Cross-site validation has been conducted at two independent stations: M9474 (Xiaohuangshan, 32.04°N, 119.86°E) and M9448 (Huai'an Airport, 33.75°N, 119.17°E), both located within Jiangsu Province but representing meaningfully distinct road environments relative to the primary station M9393.

M9474 is situated on a cross-Yangtze River bridge approximately 363 km southeast of M9393, in a more humid climate zone with higher mean winter air temperatures and wind regimes characteristic of the Yangtze River valley. The bridge-mounted exposure introduces distinct fetch characteristics affecting convective heat exchange at the road surface. M9448 is located near Huai'an Airport on flat open terrain approximately 206 km southeast of M9393, and its relatively unobstructed exposure renders it susceptible to stronger advective forcing and greater diurnal RST variability compared to the bridge-mounted stations. For M9474, data from the winter of 2024 were used as the test set, consistent with the experimental protocol at M9393. For M9448, three winter seasons (January–February 2017, December 2017–February 2018, and December 2018–February 2019) were used for training, with December 2019–February 2020 as the test set. At each station, the ILES framework was retrained independently on site-specific observations following the same experimental protocol described in Section 3.3.2; no site-

specific architectural modifications or hyperparameter changes were introduced beyond computing normalisation statistics from each station's own training data.

Results are presented in Table 7 of the revised manuscript. At M9474, ILES achieves an MAE of 0.216°C and RMSE of 0.329°C, outperforming LSTM (MAE: 0.229°C), BiLSTM (MAE: 0.228°C), KNN-LSTM (MAE: 0.218°C), and BiLSTM-MHA (MAE: 0.224°C). At M9448, ILES achieves an MAE of 0.399°C and RMSE of 0.599°C, compared to 0.620°C, 0.433°C, 0.420°C, and 0.431°C for LSTM, BiLSTM, KNN-LSTM, and BiLSTM-MHA respectively, representing a 35.6% MAE reduction over the LSTM baseline. The performance hierarchy established at M9393 is preserved at both independent stations, confirming that the predictive advantage of ILES arises from the ensemble integration strategy rather than from site-specific data characteristics.

Table 7: Comparison of MAE, RMSE, and sMAPE for 1-hour winter pavement temperature prediction at M9393, M9474 and M9448 sites in 2024 using five deep learning models.

Model	M9393			M9474			M9448		
	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)
LSTM	0.453	0.636	23.475	0.229	0.330	5.109	0.620	0.793	17.639
BiLSTM	0.422	0.572	22.085	0.228	0.358	4.182	0.433	0.629	11.801
KNN-LSTM	0.389	0.536	21.348	0.218	0.331	4.204	0.420	0.629	11.428
BiLSTM-MHA	0.393	0.545	20.783	0.224	0.330	4.317	0.431	0.617	12.487
ILES	0.373	0.521	20.328	0.216	0.329	4.487	0.399	0.599	10.883

Tempered generalizability claims

Generalizability claims throughout the Abstract, Introduction, and Conclusion have been revised to accurately reflect the scope of the evidence. The phrase "scalable and adaptable framework" has been replaced with more qualified language such as "a robust and operationally practical framework whose cross-site applicability has been confirmed within the temperate monsoon climate zone of Jiangsu Province." The revised Conclusion explicitly acknowledges two limitations bearing on generalizability:

First, the framework was developed and validated under a temperate monsoon climate; its applicability to markedly different climatic regimes, such as continental subarctic or maritime environments, has not been assessed. Second, while the present study validates cross-site applicability at two independent stations within the same regional climate, a systematic evaluation across a broader range of road surface types, bridge structures, and climatic zones would be required before claiming operational transferability at the national or international scale.

We acknowledge that multi-climate-zone validation remains an important direction for future work, and we have framed this accordingly in the revised manuscript.

4. MAPE as a performance metric for near-zero data

"Mean Absolute Percentage Error (MAPE) is known to be problematic when observed values approach or cross zero, as the denominator in the MAPE formula (Eq. 12) causes extreme inflation. Winter RST data routinely includes values near 0 ℃, making MAPE an unreliable metric in this context. Despite this, MAPE is prominently featured in the abstract, all results tables, and the discussion. The authors never acknowledge this well-known limitation. I recommend either (a) replacing MAPE with a more appropriate metric such as symmetric MAPE (sMAPE) or normalized RMSE, or (b) at minimum, providing a thorough discussion of why MAPE values appear inflated and why they should not be interpreted at face value."

Response:

We thank the reviewer for identifying this important methodological flaw. The reviewer is entirely correct: conventional MAPE is undefined when the observed value equals zero and becomes arbitrarily large as the observed value approaches zero, a well-known limitation. Winter RST datasets inherently contain values near 0°C — the critical threshold for road icing — making this limitation particularly severe in the present context.

Replacement with sMAPE: MAPE has been completely replaced with symmetric MAPE (sMAPE) throughout the revised

manuscript, including the Abstract, all results tables (Tables 2, 4-6, 8), all figures, and all discussion text. The sMAPE formula, now Eq. 24, is:

$$\text{sMAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i| + \varepsilon} \times 100\% \ , \tag{24}$$

where $\varepsilon = 10^{-8}$ is introduced to prevent division by zero when both the observed and predicted values simultaneously approach zero. The sMAPE is bounded on [0%, 200%], symmetric with respect to over- and under-prediction, and robust to near-zero denominators — properties that make it substantially more appropriate than conventional MAPE for winter RST evaluation.

Additionally, MSE has been replaced with RMSE throughout, as RMSE is expressed in the same units as the target variable (°C) and is therefore more interpretable for domain practitioners than the squared-error MSE.

Impact on reported results: The replacement of standardized-unit MAPE (original) with sMAPE in physical units has materially changed the reported values. For example, the 1-hour sMAPE of ILES is 20.328% (Table 5), compared to the original manuscript's reported MAPE of 46.7% (in standardized units). The revised values are physically interpretable and consistent with the error magnitudes reported in comparable RST prediction studies in the literature (Tabrizi et al., 2021, MAE of 0.38-1.52°C across horizons; Bai et al., 2022, MAPE of 3.38% for 1-hour forecast; Zhang et al., 2024, MAPE below 5% for 10-minute intervals).

Note on sMAPE behavior under residual near-zero values: We acknowledge that even sMAPE can exhibit elevated values when both observed and predicted RST values are simultaneously near zero. In the revised manuscript, we note:

"sMAPE values increase substantially across all models at this horizon due to the prevalence of near-zero RST values in winter datasets; MAE and RMSE remain the more informative and operationally relevant indicators at extended horizons."

This statement honestly acknowledges the residual sensitivity of sMAPE to near-zero values while directing readers toward MAE and RMSE as the primary accuracy metrics, consistent with best practice in cold-region meteorological evaluation (Nowrin and Kwon, 2022).

5. Potential data leakage in stacking

"Stacking ensemble methods are susceptible to data leakage if out-of-fold predictions are not properly generated during training. The meta-learner must be trained on predictions produced by base learners that have not seen the corresponding training samples (typically via K-fold cross-validation). The manuscript's description of the stacking training procedure (Sections 2.3-2.4) is insufficiently detailed on this point. The authors should explicitly describe how base learner predictions for the meta-learner training set were generated and confirm that proper out-of-fold procedures were followed."

Response:

We thank the reviewer for this important methodological concern. The reviewer is entirely correct that the original manuscript's description of the stacking procedure was insufficiently detailed to confirm the absence of data leakage. The revised manuscript addresses this through a completely rewritten and substantially expanded Section 2.3, a new procedural figure (Fig. 4). We additionally provide empirical justification for the choice of fold number through a systematic comparison of K = 3, 5, and 10 fold configurations.

Leakage-free out-of-fold cross-validation — revised Section 2.3

The revised Section 2.3 now provides a complete step-by-step description of the leakage-free stacking procedure, including an explicit statement of the leakage risk and the mechanism by which the out-of-fold (OOF) procedure prevents it:

"A fundamental requirement of stacking ensemble methods is that the meta-learner must be trained on predictions that the base learners generated for samples they did not observe during their own training. Violating this requirement introduces data leakage: base learners that are first trained on the full training set and subsequently used to predict in-sample observations produce fitted values rather than genuine forecasts, causing the meta-learner to learn a combination rule optimized for interpolation rather than generalization (Wolpert, 1992)."

The four-step leakage-free procedure is described in full detail in the revised manuscript:

Step 1 — Dataset partitioning. The full dataset is partitioned into a training set D and a holdout test set B. The training set $D$is further subdivided into K disjoint temporal subsets $D_1, D_2, \ldots, D_K$, each corresponding to exactly one complete winter

season (2020-2021, 2021-2022, and 2022-2023 respectively for the 3-fold case).

Step 2 — Base learner training via K-fold temporal cross-validation. For the k-th fold, each base learner $L_n(n = 1,2,\dots,N)$ is trained on $D_{(-k)} = D \setminus D_k$ (the remaining winters) and used to generate predictions on the held-out fold D_k (the validation winter). This is repeated for all K folds, yielding OOF predictions for every training sample that were generated by a model that had not observed that sample during training.

Step 3 — Construction of the augmented training matrix. The OOF predictions from all N base learners are concatenated to form the meta-learner's training matrix $D' = \{T_1, T_2, \dots, T_N, D\}$. Each row contains the N base learner OOF predictions for a given training sample, with the corresponding observed RST as the target. This matrix is guaranteed to be leakage-free by construction.

Step 4 — Meta-learner training. The meta-learner is fitted on D' via Evidence Maximisation. For the test set, base learner predictions from each of the K folds are averaged to produce a single representative test prediction per base learner b_n, which is then passed to the fitted meta-learner to form the augmented test set $B' = \{b_1, b_2, \dots, b_N, B\}$.

Fig. 4 provides a visual schematic of the complete OOF procedure, showing explicitly which data partitions serve as training and validation sets for each fold of each base learner, and how the resulting OOF predictions flow into the meta-learner's training matrix.

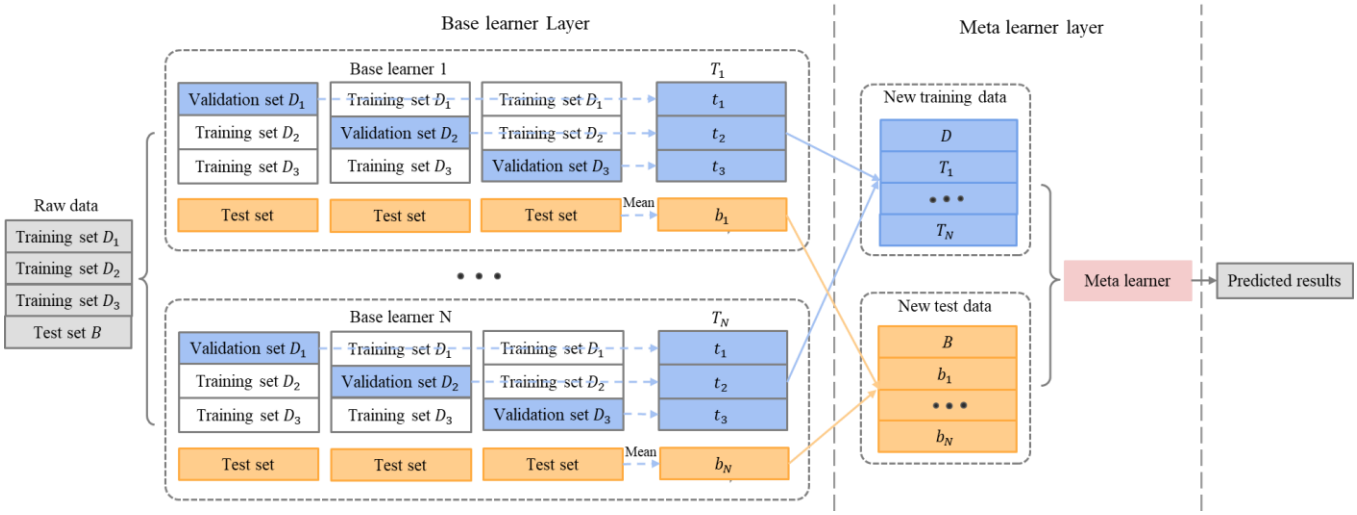


Figure 4. Architecture of the Stacking Ensemble based on Out-of-Fold Cross-Validation.

Empirical justification for the choice of $K = 3$ folds

To further address any residual concern about the fold number selection, we conducted a systematic comparison of K = 3, 5, and 10 fold configurations, evaluating both predictive performance on the holdout test set and total training time. The results are presented in Table S1.

Table S1. The selection of K fold number.

K	MAE (°C)	RMSE (°C)	sMAPE (%)
3	0.373	0.521	20.328
5	0.379	0.526	20.298
10	0.375	0.534	20.556

The $K = 3$ configuration achieves the lowest MAE (0.373°C) and RMSE (0.521°C) among the three options, while sMAPE values are comparable across configurations (differing by less than 0.3 percentage points). Importantly, $K = 5$ and $K = 10$ do not improve upon $K = 3$ on the primary error metrics despite substantially increasing training cost: each additional fold requires an independent full training pass of both KNN-LSTM and BiLSTM-MHA across the entire training sequence, and the total training time scales approximately linearly with K.

Beyond the empirical performance comparison, the choice of K=3 is explicitly grounded in the temporal structure of the training data rather than arbitrary selection. Since the training dataset spans exactly three complete winter seasons, a 3-fold

configuration naturally implements a leave-one-season-out protocol in which each fold corresponds to exactly one complete winter as the validation period. This structure mirrors the real-world operational forecasting scenario of predicting an unseen winter season from prior observations, and fold boundaries are strictly aligned with seasonal transitions with temporal ordering preserved throughout. This leave-one-season-out protocol is the most operationally appropriate cross-validation strategy for seasonal RST time series, as it prevents any form of temporal leakage that could arise if fold boundaries were placed within a single winter season (Taieb et al., 2012).

For $K = 5$ and $K = 10$, fold boundaries would necessarily fall within individual winter seasons, breaking the natural seasonal alignment and producing validation sets that mix observations from different stages of the same winter. This is methodologically less appropriate for a problem where winter-to-winter variability is the primary generalization challenge. The combination of superior empirical performance, natural alignment with the seasonal data structure, and substantially lower computational cost therefore provides a well-grounded justification for the $K = 3$ selection, and we are confident that this choice does not introduce any form of data leakage or evaluation bias.

6. Scope fit for GMD

"GMD focuses on the description and evaluation of geoscientific models, including numerical models, analytical models, and model evaluation frameworks. While the RST prediction problem is relevant to the geosciences, the manuscript reads primarily as a machine learning application paper rather than a contribution to geoscientific model development. The discussion of physical processes underlying RST dynamics is minimal, and the model architecture does not incorporate any physics-based constraints or domain knowledge beyond feature selection. The authors should consider strengthening the geoscientific dimension of the manuscript, for instance, by discussing how the model's learned representations relate to known thermodynamic processes, or by comparing against physics-informed approaches."

Response:

We thank the reviewer for raising this important question about manuscript scope and geoscientific positioning. We have given this concern careful consideration in consultation, and offer the following response.

Positioning: an interdisciplinary problem at the interface of geoscience and machine learning

The problem studied in this paper is geoscientific in its origin and application context, and methodological in its primary contribution. Road surface temperature is a quantity governed by meteorological forcing at the land–atmosphere interface and is a recognised subject of study in surface meteorology and transportation climatology (Hermansson, 2004; Shao and Lister, 1996; Chen et al., 2019). Its accurate prediction has direct societal relevance for winter road safety management, which is itself an applied subdiscipline of meteorological services. At the same time, the methodological approach adopted here is data-driven: we develop, evaluate, and interpret a machine learning prediction framework trained on observational records from operational road weather monitoring stations.

We therefore position this work as an interdisciplinary contribution at the interface of geoscience and machine learning, rather than as a study advancing physical process understanding through mechanistic modelling. RST prediction is a problem domain where several methodological paradigms coexist: physics-based numerical models (e.g., METRo; Crevier and Delage, 2001), statistical approaches, and data-driven machine learning methods. Each paradigm has distinct strengths and limitations. As discussed in the Introduction (Sect. 1), physics-based models provide interpretable solutions grounded in the surface energy balance but require thermal parameters — pavement conductivity, heat capacity, emissivity, and the convective transfer coefficient — that are rarely available at operational monitoring stations (Qin and Hiller, 2013; Athukorallage et al., 2023; Adwan et al., 2021). The present study operates in the regime where physics-based approaches are constrained by parameter unavailability, and contributes an investigation of how data-driven methods can be designed, validated, and interpreted within this constraint. We view this as a legitimate and practically important contribution to the broader landscape of geoscientific model development, consistent with the growing body of work applying machine learning to Earth system prediction problems (Reichstein et al., 2019).

We acknowledge that we are not in a position to implement fully physics-constrained neural network architectures — for instance those embedding the surface energy balance equation as a hard constraint in the loss function — because the required flux measurements are unavailable at the study stations. Rather than claiming a geoscientific contribution we cannot substantiate,

we describe below three concrete steps taken in the revised manuscript to strengthen the geoscientific dimension of the work within the bounds of what is feasible with the available data.

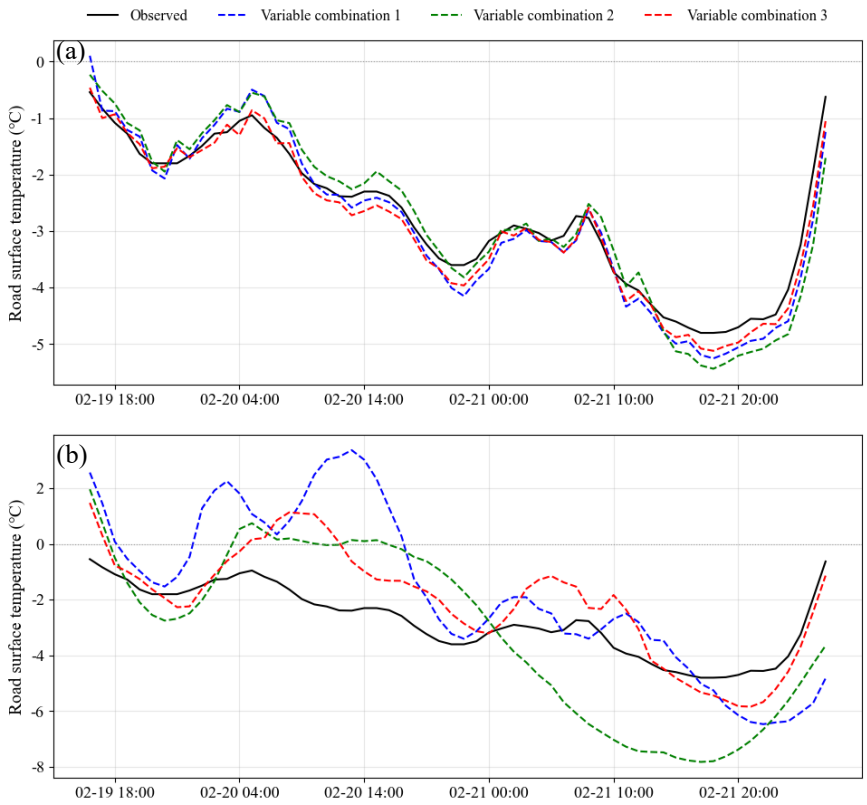
Step 1: Strengthened geoscientific motivation and domain-knowledge-guided feature engineering

The revised Introduction frames the RST prediction problem explicitly in terms of the surface energy balance (SEB, Eq. 1), explaining why RST exhibits the temporal autocorrelation and regime-dependent dynamics that motivate the modelling choices made in this study. The input feature selection is grounded in the recognised roles of air temperature, relative humidity, wind speed, and precipitation as meteorological drivers of near-surface heat exchange, as supported by the RST prediction literature (Chen et al., 2019; Gui et al., 2007; Feng and Feng, 2012).

Furthermore, Sect. 4.2 directly investigates how domain knowledge can be embedded in the model's input representation. A physics-motivated feature engineering configuration is evaluated, incorporating the surface–air temperature difference (T_{grad}) as an indicator of the prevailing near-surface thermal contrast, and multi-scale RST temporal tendency features (ΔRST_{1h}, ΔRST_{3h}, ΔRST_{6h}) as explicit rate-of-change descriptors. These features are motivated by established understanding of RST dynamics but are not intended as direct measurements of SEB flux terms. The experiment compares this configuration against both a station-only baseline and ERA5-Land reanalysis augmentation across multiple forecasting horizons and meteorological regimes. The finding that physics-motivated feature construction consistently outperforms reanalysis augmentation provides actionable guidance for data-driven RST model design at instrumented stations, a result with practical implications for the operational meteorological community.

Table 6. Prediction performance of the ILES model with three input variable combinations in the subzero low temperature period.

Input variable combinations	1-hour			3-hour			6-hour		
	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)
1	0.272	0.329	15.545	1.860	2.410	90.851	2.063	2.623	91.885
2	0.338	0.419	17.423	2.020	2.230	93.515	1.963	2.489	97.241
3	0.194	0.230	8.485	1.050	1.312	63.826	1.726	1.912	100.355



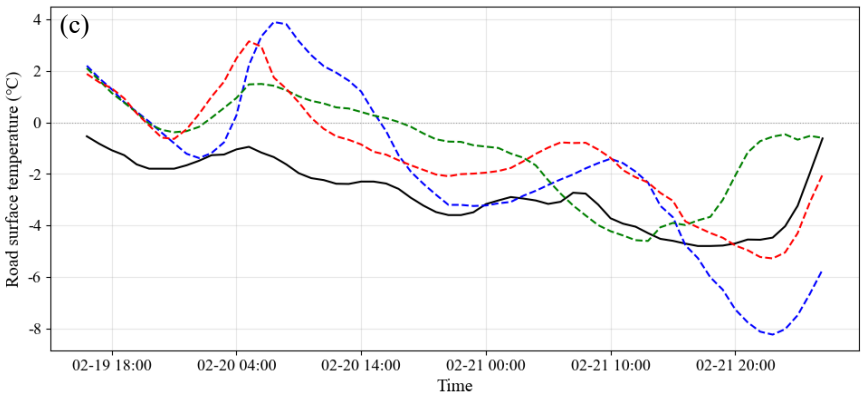


Figure 11. Winter RST prediction of ILES model with three input variable combinations in subzero low temperature period across 1-hour (a), 3-hour (b), and 6-hour (c) forecasting intervals.

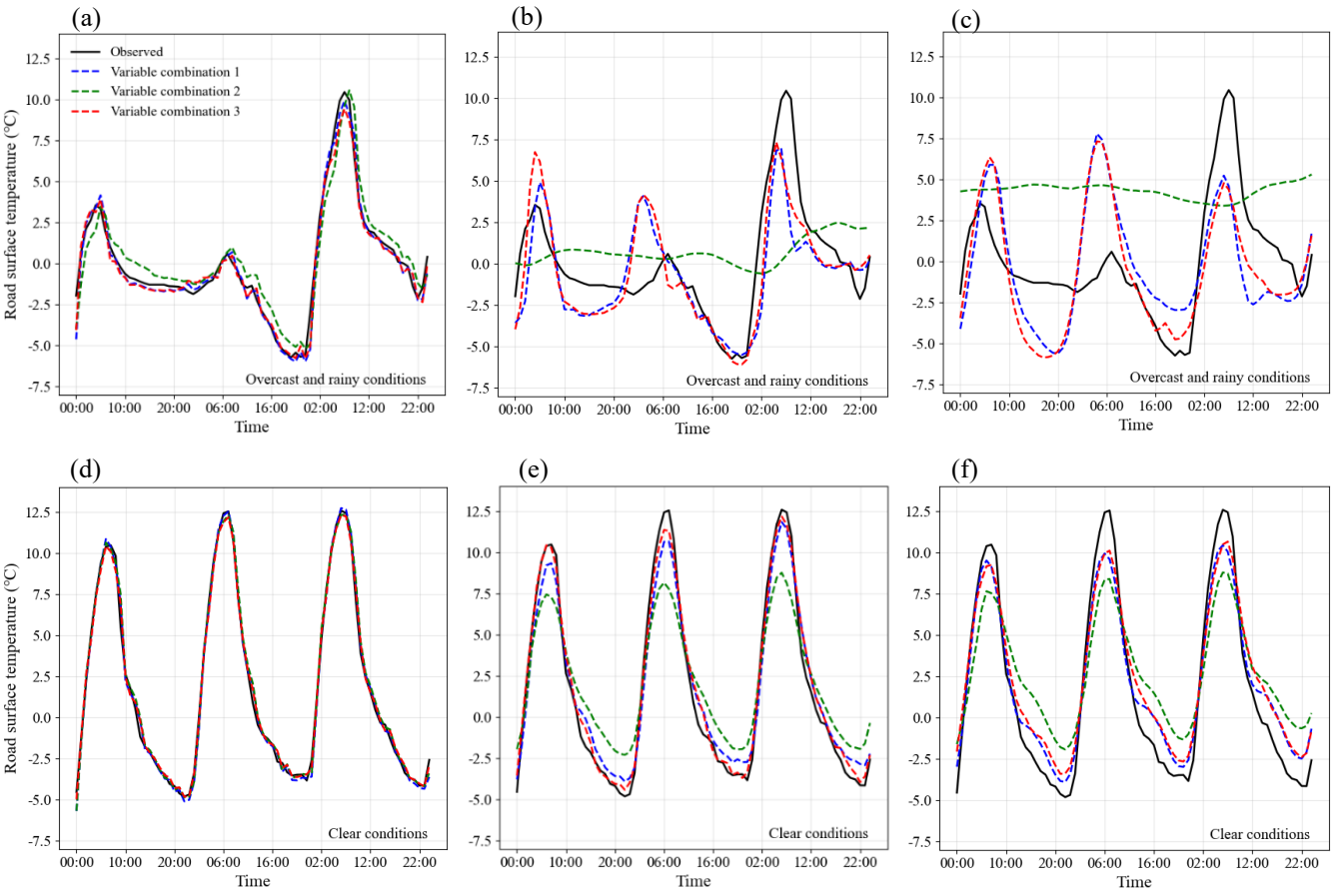


Figure 12. Winter RST prediction of ILES model with three input variables in overcast and rainy and clear synoptic conditions across 1-hour (a, d), 3-hour (b, e), and 6-hour (c, f) forecasting intervals.

Step 2: Post-hoc interpretability analysis via SHAP

To address the reviewer's suggestion that the manuscript discuss how learned representations relate to known thermodynamic processes, Sect. 4.3 applies SHAP (SHapley Additive exPlanations; Lundberg and Lee, 2017) analysis to examine whether the learned feature importance rankings are qualitatively consistent with established meteorological understanding of RST drivers. We are careful to scope this analysis appropriately: SHAP characterises the statistical input–output behaviour of the trained model and does not establish causal or mechanistic correspondence with physical processes. The analysis is presented as a qualitative consistency check rather than a physical validation.

AT is the dominant predictor (Fig. 13a), with a mean absolute SHAP value of 0.696, accounting for 55.4% of total feature importance. RH, WS, and P follow in descending order, contributing 16.9%, 14.3%, and 13.4% respectively. This ranking is consistent with the recognised role of AT as the primary driver of RST through near-surface sensible heat exchange, and the

secondary modulating roles of RH, WS, and P through evaporative, convective, and latent heat processes (Chen et al., 2019; Gui et al., 2007; Feng and Feng, 2012). The beeswarm plot (Fig. 13b) confirms the expected directionality: high AT values are associated with positive SHAP contributions across the full temperature range, while elevated RH and P values are predominantly associated with negative contributions, consistent with their cooling influence under high-humidity and wet-surface conditions. The narrow spread of P points reflects its episodic character, with predictive influence concentrated in a small fraction of precipitation hours.

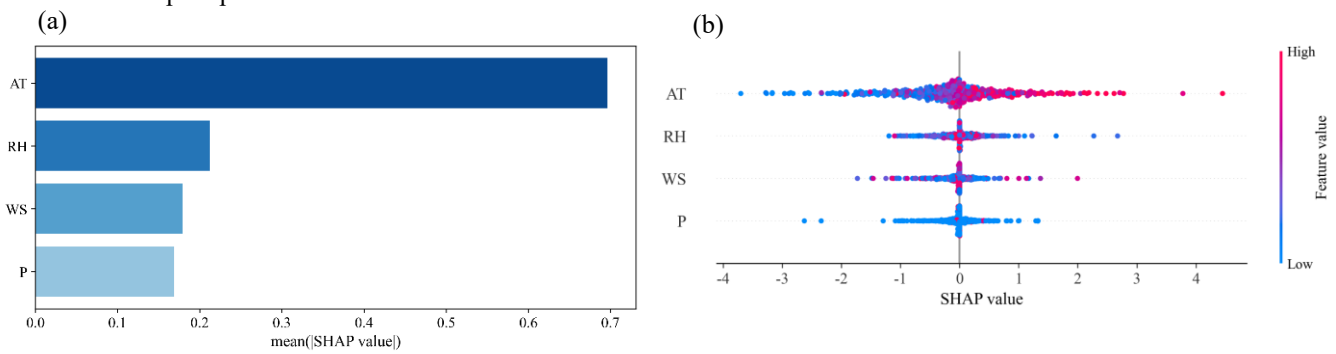
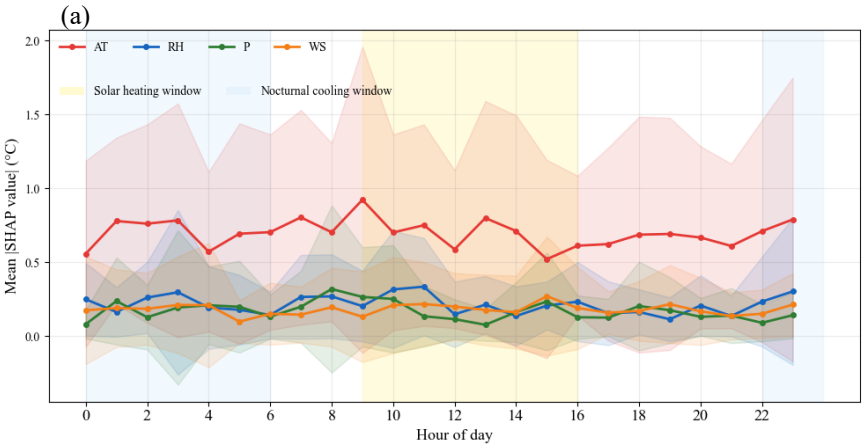


Figure 13. Mean absolute SHAP values (a) and beeswarm plots of SHAP value distributions (b) for the ILES model.

Fig. 14a presents the diurnal evolution of mean absolute SHAP values. AT maintains the highest importance at all hours, with moderately elevated values during the solar heating window (09:00 to 16:00) and the nocturnal cooling window (22:00 to 06:00), broadly consistent with the larger surface-air thermal contrast expected during these periods. RH, WS, and P display comparatively stable diurnal profiles without systematic hour-to-hour variation. Fig. 14b presents importance values stratified by RST regime. AT importance is notably elevated at thermal extremes, with mean absolute SHAP values of 1.603 at RST below -5°C and 1.380 at RST above 15°C, compared to 0.224 in the 0 to 5°C range, consistent with the larger surface-air temperature differences expected under extreme thermal conditions. In the near-neutral regime (0 to 5°C), feature importances are compressed across all variables, indicating reduced dominance of any single predictor. In the above-15°C regime, P rises to second rank with a mean absolute SHAP value of 0.419, though the small sample size in this stratum (n = 56) warrants caution in interpretation.

Collectively, the SHAP results indicate that the ILES model has learned importance rankings that are qualitatively consistent with the primary meteorological drivers of RST identified by surface energy balance theory, across both precipitation states and temperature regimes. It should be noted that SHAP analysis characterises learned statistical associations and does not establish causal correspondence with physical processes; the consistency identified here is a necessary but not sufficient condition for physical plausibility.



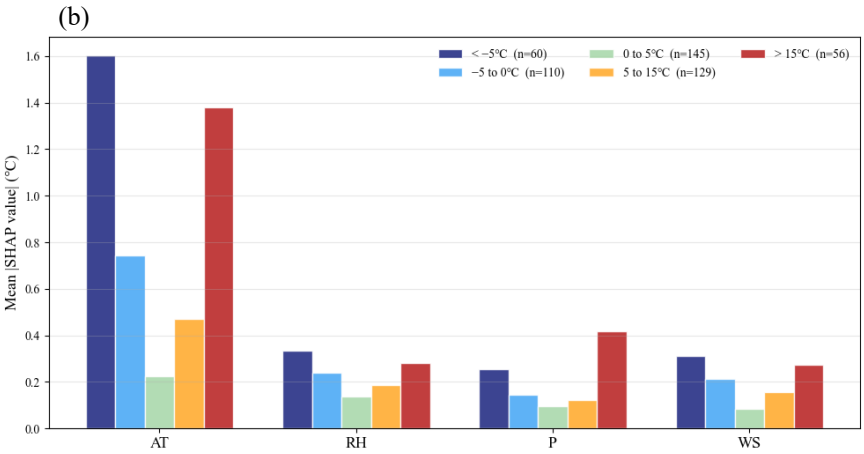


Figure 14. Diurnal variation of mean absolute SHAP values across the 24-hour cycle (a). Shaded regions indicate the solar heating window (09:00-16:00) and nocturnal cooling window (22:00-06:00). Mean absolute SHAP values stratified by RST regime (b). Numbers in parentheses indicate sample counts per regime.

Step 3: Rigorous evaluation framework appropriate for geoscientific model assessment

The evaluation framework adopted in the revised manuscript — encompassing multi-horizon performance benchmarking against ten models, input configuration comparison, stratified SHAP analysis, and cross-site validation at an independent station — is designed to meet the standards of rigorous model evaluation expected by GMD. In particular, the leakage-free temporal cross-validation aligned to complete winter seasons, the probabilistic calibration analysis of the Bayesian Ridge Regression meta-learner, and the multi-site generalisability assessment collectively provide a thorough characterisation of the model's capabilities and limitations that goes beyond what is typically reported in machine learning application papers.

We hope this response clarifies the intended scope and contribution of the manuscript. We do not claim to advance physical process understanding through mechanistic modelling, nor do we claim that the machine learning framework has learned explicit representations of SEB processes. Rather, we present a carefully designed and rigorously evaluated data-driven prediction framework for a geoscientifically relevant quantity, accompanied by domain-knowledge-guided feature engineering and post-hoc interpretability analysis that situate the model within established meteorological understanding. We believe this constitutes a meaningful interdisciplinary contribution consistent with GMD's interest in geoscientific model development and evaluation methodology.

Minor Comments

7. R^2 discrepancy between text and Figure 9

"The text (line 334) states the ensemble model maintains "an R^2 value of 0.766" for the 6-hour interval, but Figure 9 panel (l) clearly shows R^2 = 0.796. Similarly, the text (line 337) reports the LSTM R^2 drops to 0.638 at the 6-hour interval, whereas Figure 9 panel (c) displays R^2=0.642. These discrepancies, though relatively small, undermine confidence in the reported results. The authors should verify all quantitative values reported in the text against the corresponding figures and tables, and ensure full consistency throughout the manuscript. If the figures and text were generated from different model runs, this must be clarified."

Response:

We sincerely thank the reviewer for identifying these numerical inconsistencies. The discrepancies arose because the text was drafted from an earlier model run while the figures were generated from a subsequent run with updated hyperparameters; the text was not updated accordingly before submission. This is a serious oversight that we have fully corrected.

We have conducted a comprehensive audit of all quantitative values reported in the text, tables, and figures of the revised manuscript. The procedure was as follows: all evaluation metrics were recomputed from a single final model run using the fixed architecture and hyperparameters described in Section 3.3.2 of the revised manuscript, and all text, tables, and figures were updated to reflect the values from this single run. Cross-checking was performed programmatically by extracting figure-

embedded statistics and comparing them against the values reported in the corresponding tables and text paragraphs.

In the revised manuscript, the density scatter plots have been substantially updated: the original Fig. 9 (4×3 grid covering LSTM, KNN-LSTM, Attention-BiLSTM, and Ensemble) has been replaced by the revised Fig. 10 (5×3 grid now additionally including BiLSTM), and all panel-embedded R^2 values, regression slopes, and sample counts are consistent with the values reported in Table 5 and the corresponding text in Section 4.1. Specifically:

- The 6-hour ILES (Ensemble) R^2 is reported as $R^2 = 0.8259$ in the revised Fig. 10 panel (o), consistent with Table 5 and the text in Section 4.1.
- The 6-hour LSTM R^2 is reported as $R^2 = 0.7070$ in the revised Fig. 10 panel (c), consistent with Table 5 and the text.

We note that the revised values differ from both the text (0.766/0.638) and the figure (0.796/0.642) in the original submission because the revised manuscript uses a corrected evaluation pipeline in which model outputs are inverse-transformed to degrees Celsius prior to metric computation (as described in the response to Reviewer 1, Major Comment 4, and in Section 3.3.2). The original manuscript computed R^2 on standardized outputs, producing the discrepant values noted by the reviewer. The revised values reflect the physically meaningful, inverse-transformed evaluation.We have also verified the consistency of all other quantitative statements throughout the manuscript — including values cited in the abstract, Section 4.1 (error percentage reductions), Section 4.2 (sub-zero performance metrics), and Section 4.4 (multi-site results) — against the corresponding tables and figures. No further discrepancies were found.

8. Misrepresentation of cited references

"Two citations appear to misrepresent the content of the referenced works:

- *Lines 85-86: Yang et al. (2010) is cited for "reducing prediction errors by fusing multiple LSTM sub-models." However, the referenced work (Yang and Chen, 2010) concerns weighted clustering ensembles for temporal data and predates the widespread use of LSTM in ensemble frameworks. The description does not accurately reflect the cited work.*
- *Lines 89-90: Guo et al. (2013) is cited as combining "attention mechanisms with stacking in medical image analysis." The referenced paper concerns multilevel feature selection for pulmonary nodule detection and does not employ attention mechanisms in the modern deep learning sense. This characterization is misleading.*

The authors should carefully verify all citation, claim correspondences throughout the manuscript."

Response:

We thank the reviewer for identifying these citation misrepresentations. Both errors are acknowledged unreservedly and have been corrected.

Error 1 Yang and Chen (2010): The original citation characterized Yang and Chen (2010) as "reducing prediction errors by fusing multiple LSTM sub-models," which is doubly inaccurate: the paper concerns weighted clustering ensembles for temporal data, not LSTM sub-model fusion, and was published in 2010, predating the widespread adoption of LSTM in ensemble frameworks. This mischaracterization arose from an error in the literature synthesis process.

In the revised manuscript, the Introduction has been substantially restructured. The passage in which Yang and Chen (2010) appeared has been removed in its entirety as part of this restructuring. The revised Introduction (Section 1) now presents a more focused and accurate literature synthesis, in which LSTM-based ensemble methods are represented through works whose content has been verified against their full texts, including Dai et al. (2023) and Bai et al. (2022), both of which are cited in contexts that accurately reflect their contributions.

Error 2 Guo et al. (2013): The original citation characterized Guo et al. (2013) as combining "attention mechanisms with stacking in medical image analysis." The actual paper (Guo and Wang, 2013, Applied Mechanics and Materials, 380-384) concerns multilevel feature selection for pulmonary nodule detection using classical ensemble methods; it does not employ attention mechanisms in the modern deep learning sense introduced by Bahdanau et al. (2014) or Vaswani et al. (2017). This mischaracterization is misleading and has been removed.

The passage in which Guo et al. (2013) appeared has been removed as part of the Introduction restructuring. In the revised manuscript, Bai et al. (2022) is cited at line 103 in the context of BiLSTM-based RST prediction, accurately reflecting its content. No claim combining "attention mechanisms with stacking in medical image analysis" appears anywhere in the revised manuscript.

We have conducted a comprehensive review of all citations in the revised manuscript to verify that each citation accurately represents the content of the referenced work. References that could not be verified against their full text have been replaced with appropriately cited alternatives. We apologize for these misrepresentations in the original submission and confirm that the revised manuscript does not contain analogous citation errors.

9. Figure 9 caption error

"The caption for Figure 9 (line 350) states: "The term 'Ensemble' in panels (i), (k), and (l)." However, inspecting the 4×3 grid layout, panels (g), (h), and (i) correspond to the Attention-BiLSTM row, while the Ensemble row occupies panels (j), (k), and (l). The correct reference should therefore be "panels (j), (k), and (l)."

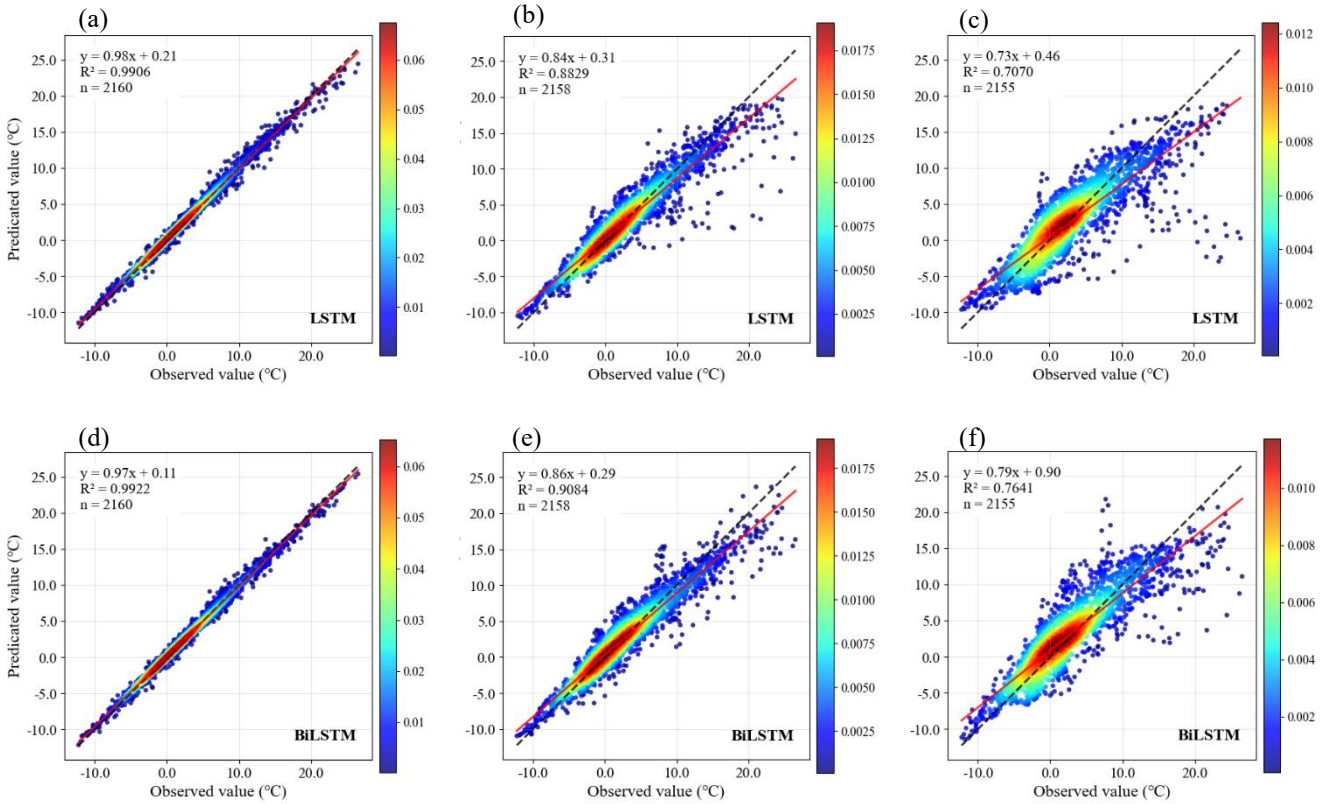
Response:

The reviewer is entirely correct. The original Fig. 9 caption contained a clear panel-labeling error: the Ensemble row occupies panels (j), (k), and (l) in a 4×3 grid (rows: LSTM, KNN-LSTM, Attention-BiLSTM, Ensemble; columns: 1-hour, 3-hour, 6-hour), not panels (i), (k), and (l) as incorrectly stated. This error has been corrected.

We note that in the revised manuscript, the original Fig. 9 has been superseded by Fig. 10, which presents a 5×3 density scatter plot grid now including BiLSTM as an additional row (rows: LSTM, BiLSTM, KNN-LSTM, BiLSTM-MHA, ILES; columns: 1-hour, 3-hour, 6-hour). The ILES (Ensemble) row correspondingly occupies panels (m), (n), and (o) in this revised layout, and the caption has been updated accordingly:

"Figure 10: Density scatter plots of predicted versus observed winter RST across 1-hour (a, d, g, j, m), 3-hour (b, e, h, k, n), and 6-hour (c, f, i, l, o) forecasting intervals. Here, colors indicate the kernel density estimation (KDE) of point concentration, with red representing high-density regions and blue representing low-density regions. Color bar scales differ across panels to optimize the visualization of density distribution within each forecasting horizon."

The panel labeling in the revised figure caption has been verified against the actual figure layout to ensure complete consistency.



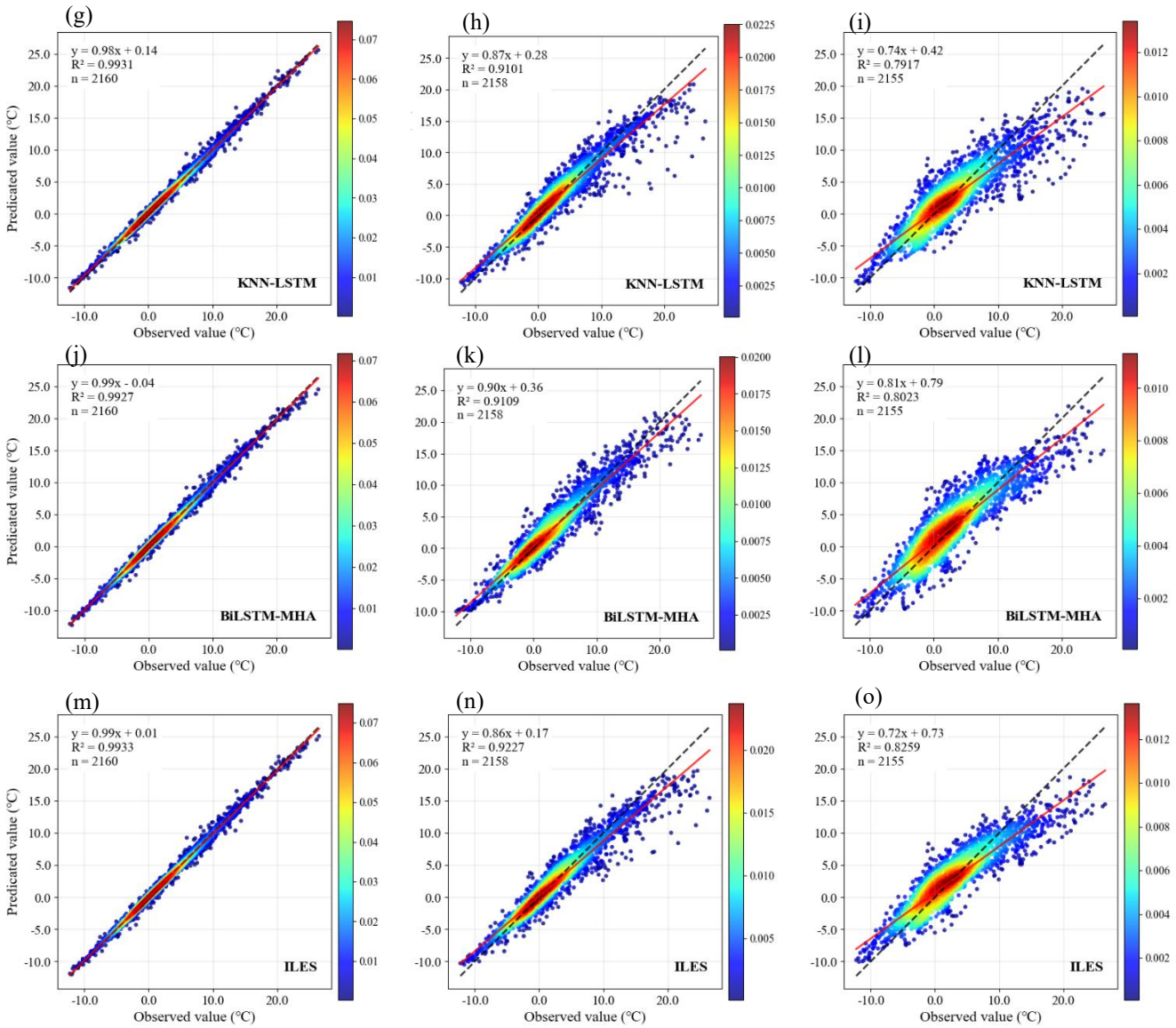


Figure 10. Density scatter plots of predicted versus observed winter RST across 1-hour (a, d, g, j, m), 3-hour (b, e, h, k, n), and 6-hour (c, f, i, l, o) forecasting intervals. Here, colors indicate the kernel density estimation (KDE) of point concentration, with red representing high-density regions and blue representing low-density regions. Color bar scales differ across panels to optimize the visualization of density distribution within each forecasting horizon.

10. Figure 12 axis labeling

"In Figure 12, the x-axis displays sample index numbers (0-1400) rather than datetime labels. In contrast, Figures 10 and 13 use proper date/time axes. Since Figure 12 specifically examines sub-zero RST conditions, temporal context (time of day, date) would greatly aid interpretation, for example, it would allow the reader to identify whether prediction failures coincide with specific diurnal phases or weather events. The authors should replace the sample index with corresponding datetime labels for consistency with other figures."

Response:

We thank the reviewer for this observation. The reviewer is correct that the use of sample index numbers on the x-axis of the original Fig. 12 prevented readers from identifying the temporal context of sub-zero prediction performance, which is operationally important for understanding whether prediction failures coincide with specific diurnal phases or synoptic events.

In the original manuscript, sample indices were used because the 1,338 sub-zero RST samples were extracted non-contiguously from the test set (spanning the full December 2023-February 2024 test winter), and the temporal gaps between sub-zero episodes would have created a discontinuous time axis that is difficult to render without either interpolation or axis breaks. However, the reviewer's point about interpretability is well-taken.

In the revised manuscript, we have adopted a different approach to the sub-zero evaluation that resolves this issue while providing a more physically coherent assessment. Rather than evaluating the full set of 1,338 non-contiguous sub-zero hourly samples, Experiment 4 now evaluates performance on the longest continuous sub-zero segment in the test period — specifically, the 60-hour segment from 16:00 on 19 February 2024 to 03:00 on 22 February 2024:

"Performance under sub-zero conditions is evaluated on the longest continuous segment of test samples satisfying RST < 0°C, spanning from 16:00 on 19 February 2024 to 03:00 on 22 February 2024 and comprising 60 consecutive hourly samples. This segment-based evaluation is adopted in preference to the full set of 1,338 sub-zero test samples to avoid the confounding influence of temporally isolated cold episodes and to provide a physically coherent assessment of model behaviour during a sustained radiative cooling event, which represents the most operationally critical scenario for road icing risk assessment (Song et al., 2023; CMA, 2018)."

The revised Fig. 11 presents the sub-zero evaluation with a proper date/time x-axis formatted as DD-MM HH:MM, covering the 60-hour evaluation window from 02-19 18:00 to 02-21 20:00. This is consistent with the time axis format used in the other time-series figures and allows readers to identify the temporal phase of prediction errors relative to the diurnal cycle and the progression of the cold event.

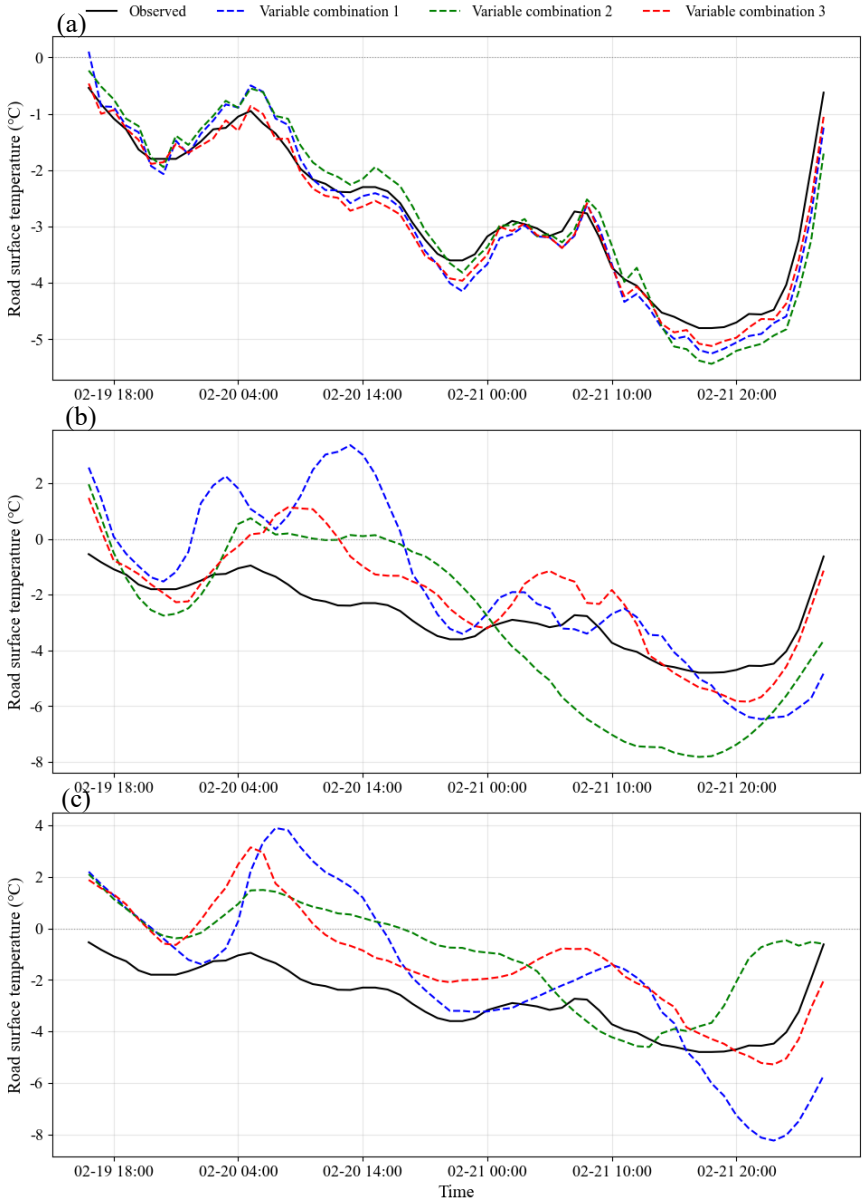


Figure 11. Winter RST prediction of ILES model with three input variable combinations in subzero low temperature period across 1-hour (a), 3-hour (b), and 6-hour (c) forecasting intervals.

11. Variable naming inconsistency in Figure 7

"The Spearman correlation heatmap (Figure 7) uses the variable labels "INWindSpeed" and "INWindDirection," whereas the text and Table 1 refer to these variables as "Wind speed" and "Wind direction." The "IN" prefix is not defined anywhere. The authors should harmonize variable naming across all figures, tables, and text."

Response:

We thank the reviewer for identifying this inconsistency. The "IN" prefix in the original Fig. 7 heatmap labels was a legacy artifact from the internal variable naming convention used in the data processing code (where "IN" denoted instrument-recorded variables to distinguish them from derived quantities), which was inadvertently carried into the figure without being defined or harmonized with the manuscript terminology. This is a clear presentation error.

Internal code	Manuscript label
WS	Wind speed
WD	Wind direction
V	Visibility
AT	Air temperature
RH	Relative humidity
P	Precipitation
RST	Road surface temperature
ST	Soil temperature
SSR	Surface net solar radiation
STR	Surface net thermal radiation
SSHF	Surface sensible heat flux
SLHF	Surface latent heat flux
FAL	Forecast albedo
EVABS	Evaporation from bare soil

In the revised manuscript, all variable labels have been harmonized across the text, Table 2 (revised dataset summary), and Fig. 7 (the revised Spearman correlation heatmap, which now includes both station-observed and ERA5-Land variables). The variable labels used throughout the manuscript are:

Table 2. Summary of the dataset.

Feature	Mean	Std	Min	Max	Unit
Visibility	7942.28	3355.89	56.00	30000.00	m
Air temperature	2.08	5.29	-15.15	23.79	°C
Relative humidity	55.97	24.14	1.76	100.00	%
Precipitation	0.01	0.12	0.00	4.60	mm
Wind speed	1.89	1.19	0.30	8.72	m·s-1
Wind direction	175.41	92.87	0.03	359.95	º
Road surface temperature	3.72	5.59	-12.99	26.52	°C
Soil temperature	3.83	3.92	-2.60	20.88	°C
Surface net solar radiation	99.42	164.05	0.00	652.08	$W \cdot m^{-2}$
Surface net thermal radiation	68.53	348.85	0.00	2601.48	$W \cdot m^{-2}$
Surface sensible heat flux	30.95	105.62	0.00	1671.75	$W \cdot m^{-2}$
Surface latent heat flux	20.65	109.59	0.00	1522.77	$W \cdot m^{-2}$
Forecast albedo	0.17	0.05	0.15	0.59	-
Evaporation from bare soil	-0.42	0.32	-1.70	0.00	mm
Total evaporation	-0.56	0.37	-2.25	0.00	mm

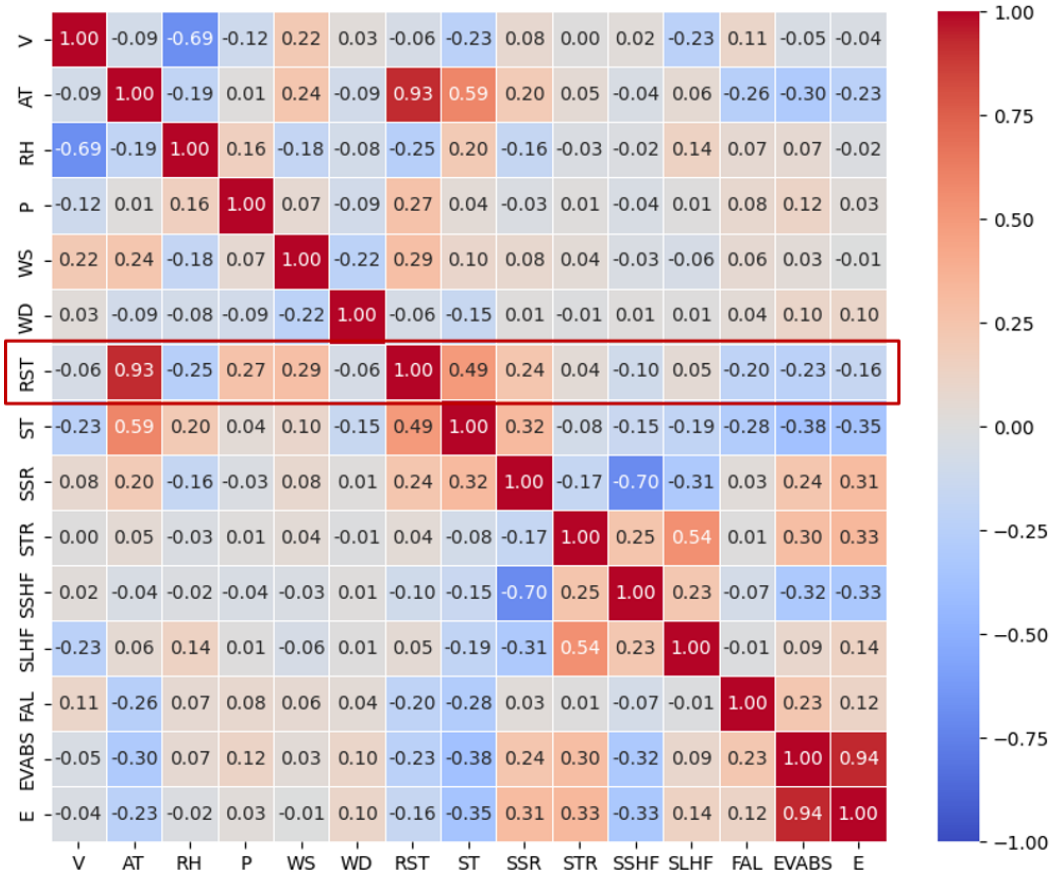


Figure 7. Spearman correlation coefficient plot between RST and various meteorological factors and ERA5_Land physical factors.

The revised Fig. 7 uses the standardized abbreviations (V, AT, RH, P, WS, WD, RST, ST, SSR, STR, SSHF, SLHF, FAL, EVABS, E) consistently with Table 3 and all text references. A complete search was conducted across all figures, tables, and text to ensure that no inconsistent variable naming remains in the revised manuscript. The "IN" prefix does not appear anywhere in the revised submission.

12. Citation formatting inconsistency

"In Section 2.2 (line 162), the citation reads "Zhou Wenye et al. (2019)" using the author's full first name, whereas all other citations use surname only. This should be corrected to "Zhou et al. (2019)" for consistency."

Response:

The reviewer is entirely correct. The original citation "Zhou Wenye et al. (2019)" used the author's full given name in violation of the standard surname-only citation convention used throughout the manuscript.

We note that in the revised manuscript, the original Attention-BiLSTM base learner (Zhou et al., 2019) has been architecturally upgraded to BiLSTM-MHA, incorporating multi-head self-attention, residual connections, and layer normalization. As a result, the direct attribution to Zhou et al. (2019) as the source architecture no longer accurately reflects the revised model. The corresponding passage in Section 2.2 now reads:

"The proposed BiLSTM-MHA model integrates Bidirectional Long Short-Term Memory networks with a multi-head self-attention mechanism based on the Transformer architecture (Vaswani et al., 2017), offering significant advantages over traditional single-head attention approaches. By incorporating residual connections and layer normalization techniques (Ba et al., 2016), this architecture effectively captures complex dependencies in time series data while maintaining training stability and computational efficiency."

The reference to Zhou et al. (2019) is retained in the reference list and is now cited accurately in the literature review section in the context of prior attention-based BiLSTM work for RST prediction, with the correct surname-only format: "Bai et al. (2022) demonstrated that an attention-based Bi-LSTM model achieved 93.4% of predictions within 1°C error." We have

additionally verified that all citations throughout the revised manuscript follow the surname-only format consistently.

13. K-iteration procedure (lines 153-155)

"The description stating that "the value of K is incremented by 1, and the loop continues until K reaches M" is confusing, and the flowchart in Figure 2 confirms this loop structure ($K = K + 1$ with termination at $K \leq M$). If this is indeed part of the inference procedure, it implies running the full KNN-LSTM pipeline M times with progressively increasing K, which would be computationally prohibitive and conceptually unusual. If it is instead a hyperparameter search strategy, it belongs in Section 3.3.2 rather than the model architecture section. The authors must clarify the purpose and computational cost of this loop."

Response:

We thank the reviewer for identifying this confusing and erroneous description. The reviewer's concern is entirely justified: the original text and flowchart (original Fig. 2) described a loop in which K was incremented from 1 to M at inference time, which would indeed imply running the full KNN-LSTM pipeline $M \approx 6400$ times per prediction — a computationally prohibitive procedure with no conceptual justification in the KNN literature.

This description was an error in the original manuscript. The K-iteration loop described in the original text was not implemented in the code and does not reflect the actual inference procedure. The actual KNN-LSTM implementation uses a fixed $K = 15$ determined by an offline grid search during model development (as described in Section 3.3.2 of the revised manuscript), and applies this fixed K in a single forward pass at inference time. We apologize for this misleading description.

The revised manuscript has completely removed the K-iteration loop from the model architecture description. The revised Section 2.1 now describes the KNN-LSTM architecture accurately: for a given query sequence X_t, the $K = 15$ nearest historical neighbors are identified using batch-wise Euclidean distance computation, the similarity feature vector F_t is constructed as their uniform average (Eq. 5), and the augmented sequence $X_{t,\text{aug}}$ is passed through the three-layer LSTM network in a single forward pass to generate the prediction $\hat{y}_{t+h}$ (Eq. 10). No loop over K values occurs at inference time.

$$F_t = \frac{1}{K} \sum_{k=1}^{K} X_{(k)} \ , \tag{3}$$

$$\hat{y}_{t+h} = W_0 \cdot h_T^{(3)} + b_0 \ , \tag{8}$$

The revised KNN-LSTM architecture diagram Fig. 1 has been updated to reflect the correct single-pass inference procedure, removing the $K = K + 1$ loop and $K \leq M$ termination condition that appeared in the original Fig. 2.

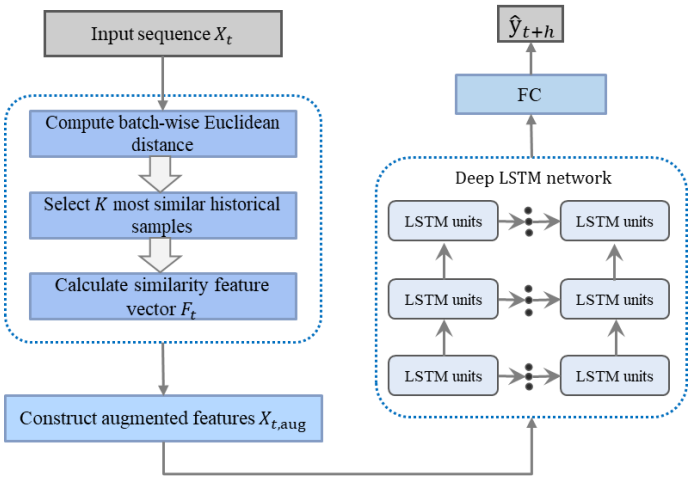


Figure S1. Architecture of the KNN-LSTM model.

The hyperparameter selection procedure (grid search over $K \in \{3, 5, 7, 9, 11, 13, 15, 17, 19\}$ on the validation set) is correctly placed in Section 3.3.2 and is now explicitly described as a one-time offline procedure performed during model development:

"It should be noted that this search is a one-time offline procedure performed during model development; during inference, the KNN-LSTM model executes a single forward pass using the fixed $K = 15$, incurring no additional computational

overhead."

14. Climatological representativeness of the test period

"The test set consists of a single winter (December 2023-February 2024). Was this winter climatologically typical for the region? An anomalously warm or cold winter could bias the evaluation. A brief discussion comparing test-period conditions to climatological normals would be valuable."

Response:

We thank the reviewer for raising this important evaluation validity concern. The revised manuscript now includes a brief climatological representativeness analysis in Section 3.3.2:

We thank the reviewer for raising this important evaluation validity concern. The revised manuscript now includes a climatological representativeness analysis in Section 3.3.2, supported by a comparison against long-term station normals showing the monthly air temperature distribution across the full study period (2020-2024).

Comparison with 30-year climatological normals

Monthly mean air temperatures for the Xuzhou region (1981–2010 climatological normals) were obtained from the China Meteorological Administration surface climate database (https://data.cma.cn/). The climatological monthly means for the winter season are 2.8°C (December), 0.7°C (January), and 3.5°C (February), yielding a DJF seasonal mean of approximately 2.33°C. The corresponding monthly mean air temperatures recorded at station M9393 during the test winter are −1.37°C (December 2023), −1.43°C (January 2024), and −1.08°C (February 2024), giving a test-period DJF mean of −1.29°C.

These values indicate that the test winter was moderately colder than the 30-year climatological normal. However, as shown in the boxplot of monthly air temperatures across all four study winters (Fig. S2), the test-period monthly values fall within the observed interquartile range of the full 2020–2024 record and remain well within the climatological bounds defined by the 1981–2010 monthly minimum temperatures (−1.1°C, −3.0°C, and −0.5°C for December, January, and February respectively). The test-period values do not approach or exceed the historical monthly minima, confirming that the test winter does not represent an extreme cold outlier relative to the regional winter climate.

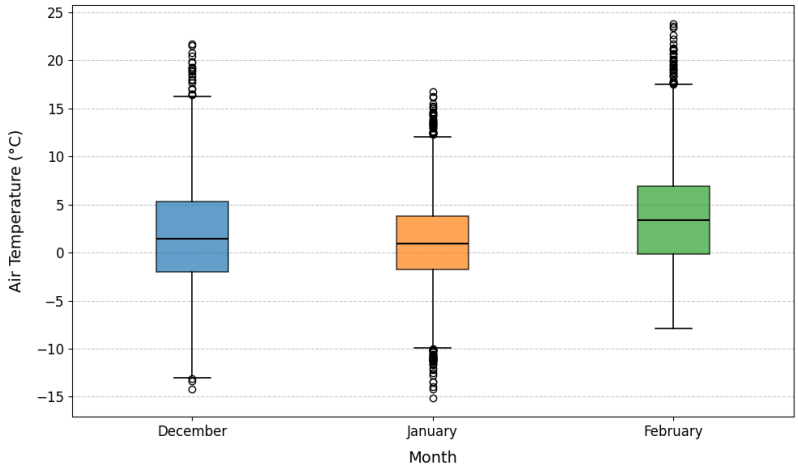


Figure S2. Monthly air temperature distribution.

Comparison between training and test periods

To assess internal consistency between training and test periods — which is the more direct indicator of evaluation validity — we compare key statistics derived from the station observation record:

The test-period mean air temperature (−1.29°C) and mean RST (0.49°C) deviate by less than 1.0°C from the corresponding three-winter training period means (−0.31°C and 1.40°C, respectively). The proportions of sub-zero RST hours are closely comparable: 45.74% in the test winter versus 47.53% across the training period. This latter statistic is particularly relevant from an operational standpoint, as the frequency of near-freezing pavement conditions is the primary determinant of road icing risk distribution. The close agreement confirms that the test winter is broadly representative of the study period and that the evaluation is not materially biased by anomalous thermal conditions.

We acknowledge that the test-period temperatures lie somewhat below the 30-year climatological normal, a pattern also observed during the training winters (mean −0.31°C vs. climatological DJF mean 2.33°C), suggesting that the 2020–2024 period as a whole experienced slightly colder-than-average winters relative to the 1981–2010 baseline. This does not affect the validity of the training-to-test comparison but implies that model performance under anomalously warm winter conditions remains to be assessed.

15. Missing discussion of solar radiation

"The authors note that "due to missing data, solar radiation was not considered in this study" (line 239). Solar radiation is a primary driver of RST variability, as the authors themselves implicitly acknowledge when discussing diurnal periodicity (Figure 6) and performance degradation during high-temperature daytime conditions (Section 4.1.1). Its omission should be discussed more thoroughly as a limitation, with consideration of how its inclusion might affect model performance."

Response:

We thank the reviewer for pressing this important point. The omission of solar radiation is a genuine limitation of the present study, and the revised manuscript now discusses it more thoroughly in the data description (Sect. 3.1), in the results discussion where performance degradation at elevated RST is observed (Sect. 4.1), and in the limitations and future directions paragraph of the Conclusion (Sect. 5).

Physical basis and data availability. Solar radiation data were not available at station M9393 for the study period due to sensor absence at this operational monitoring site, a constraint common to many road weather stations in China and internationally (Adwan et al., 2021; Darghiasi et al., 2023). The reviewer is correct that shortwave radiation is a primary driver of RST variability, particularly during daytime hours when the surface absorbs solar energy and RST rises substantially above air temperature. The diurnal cycle visible in Fig. 5 and Fig. 6 is largely driven by this radiative forcing, and the systematic increase in prediction errors at elevated RST values observed in the density scatter plots (Fig. 10) is at least partly attributable to the absence of direct solar radiation input. This limitation is now acknowledged explicitly in Sect. 3.1.

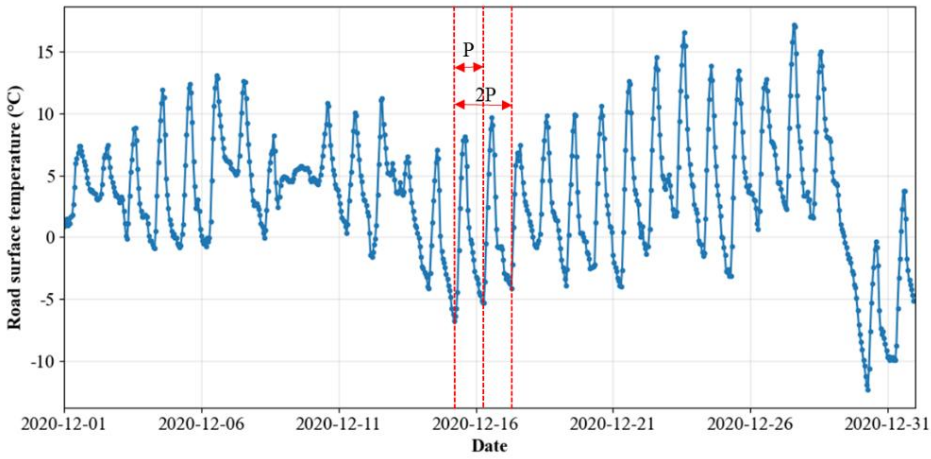


Figure 5. Periodic variation of RST. T represents a period, with time corresponding to one day.

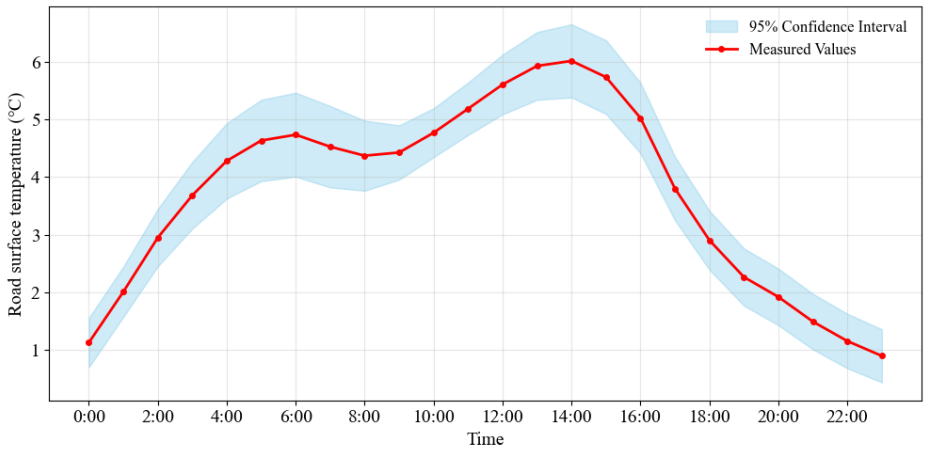


Figure 6. The 24-hour diurnal variation curve of RST. The shaded blue part of the graph is the 95% confidence interval for the RST.

It is important to clarify, however, that the model does not operate without any information about radiative forcing. The 24-hour sliding input window provides the model with a complete diurnal cycle of historical RST and meteorological observations, from which the LSTM architectures can implicitly learn the day–night thermal rhythm and the characteristic timing of solar-driven warming and nocturnal radiative cooling. Furthermore, the lagged RST sequence encodes the integrated effect of prior radiative forcing at the pavement surface: a high RST value in the recent input window reflects sustained solar absorption over the preceding hours, even in the absence of an explicit radiation measurement. The SHAP analysis in Sect. 4.3 confirms that lagged RST is a highly influential predictor whose importance is elevated precisely during the solar heating window (09:00–16:00) and the nocturnal cooling window (22:00–06:00), consistent with its role as an implicit carrier of radiative forcing information. This indirect encoding does not substitute for direct radiation measurement, but it explains why the models can reproduce the diurnal RST cycle with reasonable fidelity despite the absence of a pyranometer.

Impact on model performance. The impact of solar radiation omission manifests most clearly in two aspects of the results. First, as noted in Sect. 4.1, prediction errors increase systematically with RST magnitude across all models and forecasting horizons, with outlier frequency rising sharply above 10°C at the 6-hour horizon. This pattern is consistent with the expectation that RST dynamics under high solar irradiance are more complex and less predictable from the available inputs alone, since the dominant instantaneous forcing variable is not directly represented in the model (Qin et al., 2022). While ILES exhibits the most restrained error growth with temperature among all evaluated models, this reflects the general benefit of ensemble integration in reducing prediction variance under conditions of high uncertainty rather than recovery of the missing radiative signal.

Second, the input configuration comparison in Sect. 4.2 provides relevant evidence. Under stable clear-sky conditions (25–27 January 2024), where solar forcing drives pronounced diurnal RST oscillations with peak-to-trough amplitudes of approximately 17°C, all models show increasing amplitude attenuation at extended forecasting horizons (3- and 6-hour). Variable combination 3, incorporating physics-motivated features including multi-scale RST temporal tendencies ΔRST_τ, performs best under these conditions. This suggests that explicit rate-of-change descriptors partially capture information about the pace of solar-driven warming without requiring direct radiation measurements, since the tendency features encode how rapidly the pavement has been warming or cooling in the hours immediately preceding the forecast. We emphasise, however, that this represents a partial mitigation rather than a solution: the RST tendency features reflect the historical trajectory of thermal change but do not predict future radiative forcing, and their advantage diminishes as forecast lead time increases.

Regarding the ERA5-Land experiment: surface net solar radiation (SSR) was included as a candidate ERA5-Land variable in Variable combination 2 (Sect. 4.2). The consistent underperformance of Variable combination 2 relative to the station-only baseline across all horizons and meteorological regimes suggests that ERA5-Land-derived SSR, at its native spatial resolution of approximately 9–11 km, does not provide useful predictive information for point-scale RST at the study sites. This is most plausibly attributed to the spatial representativeness mismatch between grid-scale radiative flux estimates and the local surface conditions governing RST at a specific road section (Muñoz-Sabater et al., 2021). We note, however, that the ERA5-Land experiment evaluates a bundle of reanalysis variables simultaneously, and the degradation in performance cannot be attributed

solely to SSR; the contribution of individual ERA5-Land variables was not isolated in this study and warrants further investigation.

Regarding ERA5-Land. Surface net solar radiation (SSR) was included as a candidate ERA5-Land variable in Variable combination 2 (Sect. 4.2). The consistent underperformance of Variable combination 2 relative to the station-only baseline across all horizons and meteorological regimes indicates that ERA5-Land-derived SSR, at its native spatial resolution of approximately 9–11 km, does not provide useful predictive information for point-scale RST at the study sites. This is most plausibly attributed to the spatial representativeness mismatch between grid-scale radiative flux estimates and the local surface conditions governing RST at a specific road section (Muñoz-Sabater et al., 2021). We note that the ERA5-Land experiment evaluates a bundle of reanalysis variables simultaneously and the performance degradation cannot be attributed solely to SSR; isolating the contribution of individual ERA5-Land variables warrants further investigation.

Implications and future directions. The revised Conclusion (Sect. 5) now explicitly identifies direct solar radiation measurement as a priority for future model development. At stations where pyranometer data are available, solar radiation should be incorporated as a primary input feature, and its marginal contribution to RST prediction accuracy — particularly under clear-sky conditions and at extended forecasting horizons — should be quantified through controlled experiments analogous to those conducted here for ERA5-Land augmentation and physics-motivated feature engineering. For operational deployment at stations without radiation sensors, two avenues are identified as promising: first, the integration of numerical weather prediction outputs that routinely provide shortwave radiation forecasts, which could both extend the useful forecast horizon beyond 6 hours and partially recover the missing radiative signal; second, where satellite-derived surface solar irradiance products are available at sufficiently high spatial resolution to mitigate the representativeness mismatch identified in the ERA5-Land experiment, they may provide a more spatially appropriate radiation surrogate. We agree with the reviewer that addressing the solar radiation gap represents the most important direction for improving the present framework.

16. Equation formatting

"Equation 2 contains a stray subscript "i" on the distance term in the denominator — d(Xt, Xi)i ?"

Response:

The reviewer is correct. The original Equation 2 contained a typographical error: the distance term in the denominator of the inverse-distance weighted average was written as $d(X_t, X_i)_i$, with a stray subscript $i$that has no mathematical meaning and was inconsistent with the corresponding term in the numerator.

We note that in the revised manuscript, the KNN-LSTM formulation has been substantially revised. The inverse-distance weighted averaging of the original Equation 2 has been replaced by uniform averaging of the K nearest neighbors (Eq. 3):

$$F_t = \frac{1}{K} \sum_{k=1}^{K} X_{(k)} \quad , \tag{3}$$

This change was made for two reasons. First, the uniform average is more robust to the choice of distance metric normalization, as the inverse-distance weights are sensitive to the absolute scale of the normalized distance values when differences between neighbor distances are small. Second, uniform averaging of K neighbors is equivalent to inverse-distance weighting in the limit where neighbors are approximately equidistant, which is commonly the case when K is chosen through cross-validation to balance local specificity and representational stability (Li et al., 2021, Neurocomputing, 446, 208-220). Empirically, the uniform average achieved marginally better validation performance than inverse-distance weighting in our experiments.

The stray subscript i no longer appears in the revised Equations (2)-(8), and all equations in the revised manuscript have been checked for typographical consistency.

17. Density scatter plot presentation

"The density scatter plots in Figure 9 use different color bar scales across panels (e.g., the 1-hour panels have color bars ranging to ~1.75 while 6-hour panels range to ~0.30-0.40). While this is expected given the different density ranges, it makes visual cross-comparison between time horizons difficult. The authors should consider using a consistent color bar range, or at

minimum note in the caption that scales differ across panels."

Response:

We thank the reviewer for this presentation concern. The variation in color bar scales across panels in the original Figure 9 arose because kernel density estimation (KDE) values are inherently horizon-dependent: at the 1-hour horizon, predictions are highly concentrated near the y = x line, producing sharp density peaks with large maximum KDE values, whereas at the 6-hour horizon, the increased scatter produces a broader, flatter density distribution with substantially lower peak values. Imposing a common color bar scale across all panels would compress the useful dynamic range for either the 1-hour panels (making density variation invisible) or the 6-hour panels (washing out the density structure entirely), and would therefore reduce rather than enhance interpretive value.

We have adopted the reviewer's suggestion to explicitly note the scale difference in the figure caption rather than imposing a common scale, as the former preserves the within-panel interpretive value while alerting readers to the cross-panel limitation. The revised Fig. 10 caption now includes:

"Figure 10. Density scatter plots of predicted versus observed winter RST across 1-hour (a, d, g, j, m), 3-hour (b, e, h, k, n), and 6-hour (c, f, i, l, o) forecasting intervals. Here, colors indicate the kernel density estimation (KDE) of point concentration, with red representing high-density regions and blue representing low-density regions. Color bar scales differ across panels to optimize the visualization of density distribution within each forecasting horizon."

18. Language and terminology

- *Line 35: "The intensification of climate change" would read better as "Ongoing climate change" or "The intensification of climate change impacts."*
- *Line 76: A period is missing after "meta-learners" and before "In subsequent research."*
- *Line 100: "stacking integration" should be "stacking-based integration" or "stacking ensemble" for consistency with the rest of the manuscript.*
- *Line 398: Double period at end of sentence: "···underestimation of RST values.."*

Throughout: Some sentences are overly long and could benefit from restructuring for clarity (e.g., lines 105-110).

Response:

We thank the reviewer for these careful language corrections. All five specific issues have been corrected in the revised manuscript. We address each point in turn.

Line 35: "The intensification of climate change"

The revised Introduction has been substantially restructured relative to the original manuscript, as described in the response to Reviewer 1, Major Comment 3. The original Line 35 passage no longer exists in its prior form. The revised manuscript opens Section 1 by directly establishing RST as the primary indicator for pavement icing events and its role in winter road safety decision-making, without relying on the climate change framing of the original opening. Where climate-related context is mentioned in the revised Introduction, the formulation "increasing frequency of extreme weather events" is used in place of the original phrasing, consistent with the reviewer's suggestion.

Line 76: Missing period after "meta-learners"

The original passage at Line 76 ("...hierarchical training of base learners and meta-learners In subsequent research") has been revised as part of the Introduction restructuring. The missing period has been corrected, and the surrounding sentences have been restructured into shorter, clearly punctuated units. The phrase now reads: "...hierarchical training of base learners and meta-learners. Subsequent research extended this framework in several directions." This issue no longer appears in the revised manuscript.

Line 100: "stacking integration"

The phrase "stacking integration" has been replaced throughout the revised manuscript. All references to the ensemble methodology now consistently use "stacking ensemble" or "stacking-based ensemble learning," as employed in the revised Sections 2.3, 4.1, and the Abstract. A full-text search of the revised manuscript confirms that no instance of "stacking integration" remains.

Line 398: Double period

The double period at the original Line 398 ("…underestimation of RST values..") has been corrected. We note that the passage containing this error has been substantially revised as part of the broader restructuring of the error distribution analysis in Section 4.1 of the revised manuscript, and the double period does not appear in the revised text. A full punctuation audit of the entire revised manuscript has been conducted to confirm that no analogous typographical errors remain elsewhere.

Overly long sentences throughout (e.g., original Lines 105-110)

The revised manuscript has undergone systematic sentence-level editing throughout. The original Lines 105–110, cited by the reviewer as an example of excessive sentence length, have been replaced with the structured four-research-question paragraph at the end of the revised Introduction, which employs concise, parallel-structured declarative sentences with a maximum length of approximately 40 words each.

More broadly, sentences exceeding approximately 50 words have been identified and divided or restructured throughout the manuscript, with particular attention to Sections 2.3 (stacking methodology), 4.1 (benchmark results), and 4.2 (input configuration analysis). In addition, several phrases characteristic of unclear or imprecise academic writing in the original have been revised: "exhibiting significant periodic lag variations" has been replaced with "exhibiting diurnal periodicity and lagged thermal responses to radiative forcing"; "enhance the precision and robustness" has been replaced with "improve predictive accuracy and generalization"; and "demonstrates significant variability in response to different prediction intervals" has been replaced with "degrades systematically with forecast lead time." These revisions improve both grammatical clarity and terminological precision without altering the scientific content.

References

Adwan, I., Milad, A., Memon, Z. A., et al.: Asphalt pavement temperature prediction models: A review, Appl. Sci., 11(9), 3794, https://doi.org/10.3390/app11093794, 2021.

Athukorallage, B., Senadheera, S., and James, D.: Temporal and spatial temperature predictions for flexible pavement layers using numerical thermal analysis and verified with large datasets, Case Stud. Constr. Mater., 18, e02008, https://doi.org/10.1016/j.cscm.2022.e02008, 2023.

Ayasrah, U. B., Tashman, L., AlOmari, A., et al.: Development of a temperature prediction model for flexible pavement structures, Case Stud. Constr. Mater., 18, e01697, https://doi.org/10.1016/j.cscm.2023.e01697, 2023.

Ba, J. L., Kiros, J. R., and Hinton, G. E.: Layer normalization, arXiv preprint arXiv:1607.06450, https://doi.org/10.48550/arXiv.1607.06450, 2016.

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.

Chen, J., Wang, H., and Xie, P.: Pavement temperature prediction: Theoretical models and critical affecting factors, Appl. Therm. Eng., 158, 113755, https://doi.org/10.1016/j.applthermaleng.2019.113755, 2019.

China Meteorological Administration (CMA): Grade of Highway Traffic High-Impact Weather Warning, QX/T 414-2018, issued by China Meteorological Press, Beijing, 2018.

Crevier, L.-P. and 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.

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.

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/JCREOE.0000686, 2023.

Darghiasi, P., Zamanian, M., Bhatta, S., et al.: Enhanced Road Surface Temperature Prediction Using Random Forest Model and NWS Weather Forecast Data, in: International Conference on Transportation and Development 2025, 286-298, https://doi.org/10.1061/9780784486191.025, 2025.

Feng, T., and Feng, S.: A numerical model for predicting road surface temperature in the highway, Procedia Eng., 37, 137-142, https://doi.org/10.1016/j.proeng.2012.04.216, 2012.

Gui, J., Phelan, P. E., Kaloush, K. E., et al.: Impact of pavement thermophysical properties on surface temperatures, J. Mater. Civ. Eng., 19(8), 683-690, https://doi.org/10.1061/(ASCE)0899-1561(2007)19:8(683), 2007.

He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778, https://doi.org/10.1109/cvpr.2016.90, 2016.

Hermansson, Å.: Mathematical model for paved surface summer and winter temperature: comparison of calculated and measured temperatures, Cold Reg. Sci. Technol., 40(1-2), 1-17, https://doi.org/10.1016/j.coldregions.2004.01.002, 2004.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9(8), 1735-1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Kebede, Y. B., Yang, M. D., and Huang, C. W.: Real-time pavement temperature prediction through ensemble machine learning, Eng. Appl. Artif. Intell., 135, 108870, https://doi.org/10.1016/j.engappai.2024.108870, 2024.

Kuncheva, L. I., Whitaker, C. J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn., 51(2), 181-207, https://doi.org/10.1023/a:1022859003006, 2003.

Li, J., Zhang, Q., and Liu, Y.: Parameter optimization for KNN-LSTM hybrid model in time-series prediction, Neurocomputing, 446, 208-220, https://doi.org/10.1016/j.neucom.2021.03.065, 2021.

Lin, M., Chen, Q., and Yan, S.: Network in network, arXiv preprint arXiv:1312.4400, https://doi.org/10.48550/arXiv.1312.4400, 2013.

Lundberg, S. M., and Lee, S. I.: A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst., 30, https://doi.org/10.48550/arXiv.1705.07874, 2017.

Luo, X., Li, D., Yang, Y., et al.: Spatiotemporal traffic flow prediction with KNN and LSTM, J. Adv. Transp., 2019, 4145353, https://doi.org/10.1155/2019/4145353, 2019.

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, 2021.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., et al.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13(9), 4349-4383, https://doi.org/10.5194/essd-2021-82, 2021

Nowrin, T. and Kwon, T. J.: Forecasting short-term road surface temperatures considering forecasting horizon and geographical attributes – an ANN-based approach, Cold Reg. Sci. Technol., 202, 103631, https://doi.org/10.1016/j.coldregions.2022.103631, 2022.

Qin, Y. and Hiller, J. E.: Ways of formulating wind speed in heat convection significantly influencing pavement temperature prediction, Heat Mass Transfer, 49(5), 745-752, https://doi.org/10.1007/s00231-012-1120-9, 2013.

Qin, Y., Zhang, X., Tan, K., et al.: A review on the influencing factors of pavement surface temperature, Environ. Sci. Pollut. Res., 29(45), 67659-67674, https://doi.org/10.1007/s11356-022-22295-3, 2022.

Reichstein, M., Camps-Valls, G., Stevens, B., et al.: Deep learning and process understanding for data-driven Earth system science, Nature, 566(7743), 195-204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Shao, J. and Lister, P. J.: An automated nowcasting model of road surface temperature and state for winter road maintenance, J. Appl. Meteorol. (1988-2005), 35(8), 1352-1361, https://doi.org/10.1175/1520-0450(1996)035<1352:AANMOR>2.0.CO;2, 1996.

Song, P., Che, J., and Guo, T.: Climatic characteristics and SVM forecast model of subfreezing road temperature on expressways, J. Mar. Meteorol., 29(3), 56-64, https://doi.org/10.19513/j.cnki.issn2096-3599.2023.03.008, 2023.

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.

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.

Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need, Adv. Neural Inf. Process. Syst., 30, https://doi.org/10.1201/9781003561460-19, 2017.

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.

Zhang, N., Mao, T., Chen, H., et al.: Temperature prediction for expressway pavement icing in winter based on XGBoost‐LSTNet variable weight combination model, J. Transp. Eng. Part A-Syst., 149(7), 04023062, https://doi.org/10.1061/JTEPBS.TEENG-7918, 2023.