Dear Reviewers:

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. *"The manuscript's positioning is currently ambiguous across geoscience, Machine Learning (ML) methodology, and winter road operations. From a geoscientific perspective, the paper does not clearly advance physical understanding of the processes controlling RST. From an information engineering perspective, the modeling framework (improved LSTMs + stacking) appears to extend established components with limited methodological novelty. From a winter road management perspective, the operational meaning of the improved accuracy (e.g., how it would change decision-making or warning performance) is not yet sufficiently articulated, despite claims about operational deployment. Please state clearly what the primary contribution is (geoscientific, methodological, or operational) and align the framing and discussion accordingly."*

Response:

We sincerely thank the reviewer for this precise and constructive diagnosis. We fully accept the criticism and have substantially restructured the manuscript's framing to eliminate the ambiguity identified.

Upon careful reflection, and consistent with discussions among the author team, we agree that the **primary contribution of this paper is methodological**: the design, empirical justification, and systematic evaluation of the ILES (Improved LSTM Ensemble with Stacking) framework as a principled prediction methodology for winter RST forecasting. The revised manuscript is now framed consistently around this methodological contribution throughout the Introduction, Methodology, Results, and Conclusion sections. We clarify below precisely what the methodological contributions are, and how the manuscript has been revised accordingly.

We wish to be explicit about the scope of the paper's claims in response to the reviewer's tripartite diagnosis:

- Geoscientific: The paper does not claim to advance physical understanding of RST-controlling processes through mechanistic or process-based modeling. Physical understanding of the surface energy balance (SEB) is instead used as domain knowledge to motivate feature engineering choices and to interpret learned model behaviour via SHAP analysis — but the paper does not derive new process-level insights. This distinction is now stated explicitly in the revised Introduction and the SHAP discussion (Section 4.3).
- Operational: The paper does not primarily address how improved forecast accuracy would alter specific road maintenance decision protocols or warning thresholds. Operational relevance (e.g., the probabilistic icing-risk indicator derived from the BRR meta-learner's prediction intervals) is discussed where appropriate but is a secondary rather than primary contribution. The revised manuscript no longer makes unqualified claims about "operational deployment" without supporting evidence.
- Methodological: The paper's value rests on the methodological framework itself — specifically on the empirical evidence that the ILES framework is justified, interpretable, and generalizable — as elaborated in the four methodological work streams described below.

Methodological Work Stream 1: Architectural innovations in the base learners

The revised KNN-LSTM (Section 2.1) introduces two substantive extensions beyond the original Luo et al. (2019) architecture for traffic flow prediction: (i) batch-wise normalized Euclidean distance computation for numerical stability across heterogeneous meteorological inputs; and (ii) a three-layer deep LSTM replacing the original single-layer architecture to enable hierarchical temporal representation learning across multi-hour RST sequences. The revised BiLSTM-MHA (Section 2.2) replaces the single-head dot-product attention of prior work (e.g., Bai et al., 2022) with multi-head self-attention based on the Transformer architecture (Vaswani et al., 2017), incorporating residual connections (He et al., 2016) and layer normalization (Ba et al., 2016) for training stability. These architectural modifications are not mere reapplications of existing components;

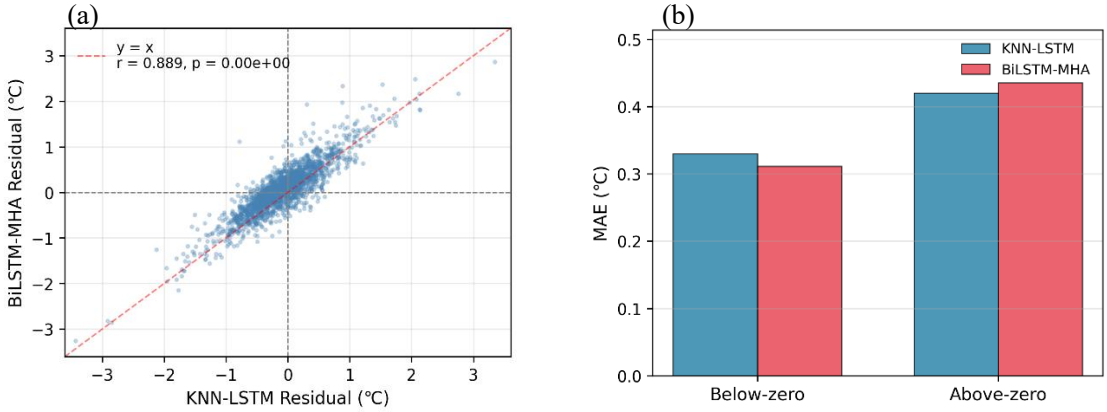
they constitute targeted technical innovations motivated by the specific challenges of RST time-series forecasting, and their empirical contributions are quantified in the ablation study (Table 4).

Methodological Work Stream 2: Empirical justification of base learner complementarity via complementarity analysis and ablation study

A key methodological weakness of the original manuscript was that the claim of "complementarity" between KNN-LSTM and Attention-BiLSTM was stated qualitatively without empirical substantiation. The revised manuscript addresses this directly through a complementarity analysis and ablation study at the 1-hour forecasting horizon (Section 3.4, Fig. 9, Table 4).

The complementarity analysis examines four dimensions: (a) residual correlation between the two base learners, indicating substantial but imperfect co-variation in prediction errors; (b) temperature-regime error decomposition, showing that BiLSTM-MHA achieves lower MAE under sub-zero conditions while KNN-LSTM performs better under above-zero conditions. This asymmetry is noteworthy because sub-zero and above-zero temperature regimes are associated with distinct meteorological forcing characteristics — sustained radiative cooling under clear skies versus variable solar heating, respectively (Hermansson, 2004) — suggesting that the two architectures may be capturing different aspects of RST dynamics; (c) comparable daytime and nighttime aggregate performance, confirming that temperature-regime complementarity is the dominant source of diversity rather than the diurnal cycle per se; and (d) sample-level advantage distribution showing BiLSTM-MHA outperforming KNN-LSTM on 50.7% of test samples and KNN-LSTM outperforming on 49.3%, confirming sustained bidirectional diversity with no consistent dominance (Kuncheva and Whitaker, 2003).

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 (0.054°C) is smaller than the arithmetic sum of the two individual gains (0.079°C), indicating that KNN-LSTM and BiLSTM-MHA partially share their areas of improvement; nonetheless, the full ILES framework achieves the best overall performance, confirming that the two components provide complementary rather than redundant contributions.



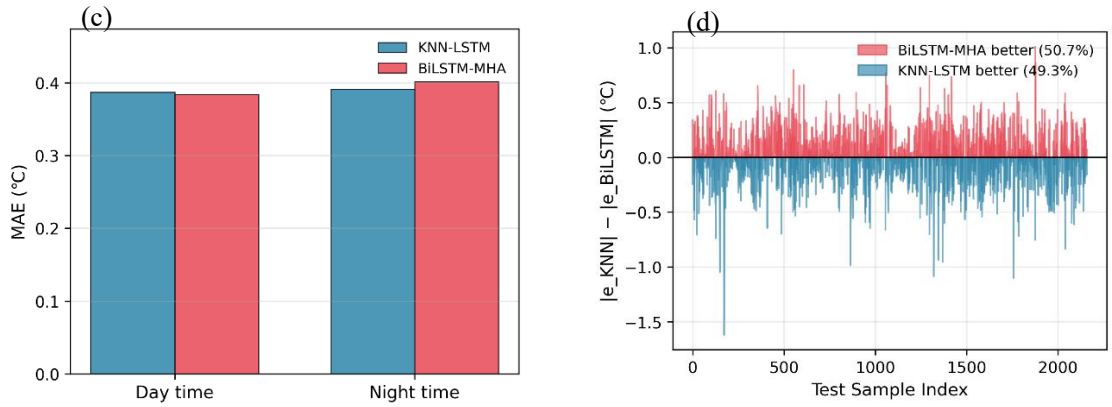


Figure 9. Complementarity analysis of base learners. Where (a) residual correlation analysis, (b) temperature interval error decomposition, (c) day and night time error decomposition, (d) time series advantage distribution.

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

Methodological Work Stream 3: Physics-motivated feature engineering as an input design strategy

An important methodological question that was absent from the original manuscript concerns how to construct informative inputs for data-driven RST models at instrumented sites where direct radiative forcing measurements (e.g., shortwave radiation) are unavailable. The revised manuscript addresses this through a controlled comparison of three input configurations (Section 4.2): a station-only baseline; ERA5-Land reanalysis augmentation; and physics-motivated feature engineering incorporating the surface–air temperature difference (T_{grad}) as a proxy for the direction of near-surface heat exchange, and multi-scale RST temporal tendency features ($\Delta\mathrm{RST}_{\tau}$ for $\tau \in \{1,3,6\}$-hour) as explicit rate-of-change descriptors constructed directly from existing station observations.

This comparison is methodologically motivated by the observation that RST tendency features explicitly encode the integrated effect of multi-hour meteorological forcing — reducing the burden on recurrent layers to extract such patterns implicitly — and that T_{grad} captures the thermal contrast relevant to icing risk without requiring additional sensors. The empirical finding that physics-motivated feature engineering consistently outperforms ERA5-Land augmentation across all forecasting horizons and meteorological regimes (Table 6, Figure 11-12) constitutes a methodologically actionable result: predictive performance at instrumented sites can be improved through domain-knowledge-guided feature construction alone, without recourse to external reanalysis products. This finding directly addresses the question of input design strategy for operational RST forecasting, a methodological gap not addressed in the prior RST prediction literature.

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

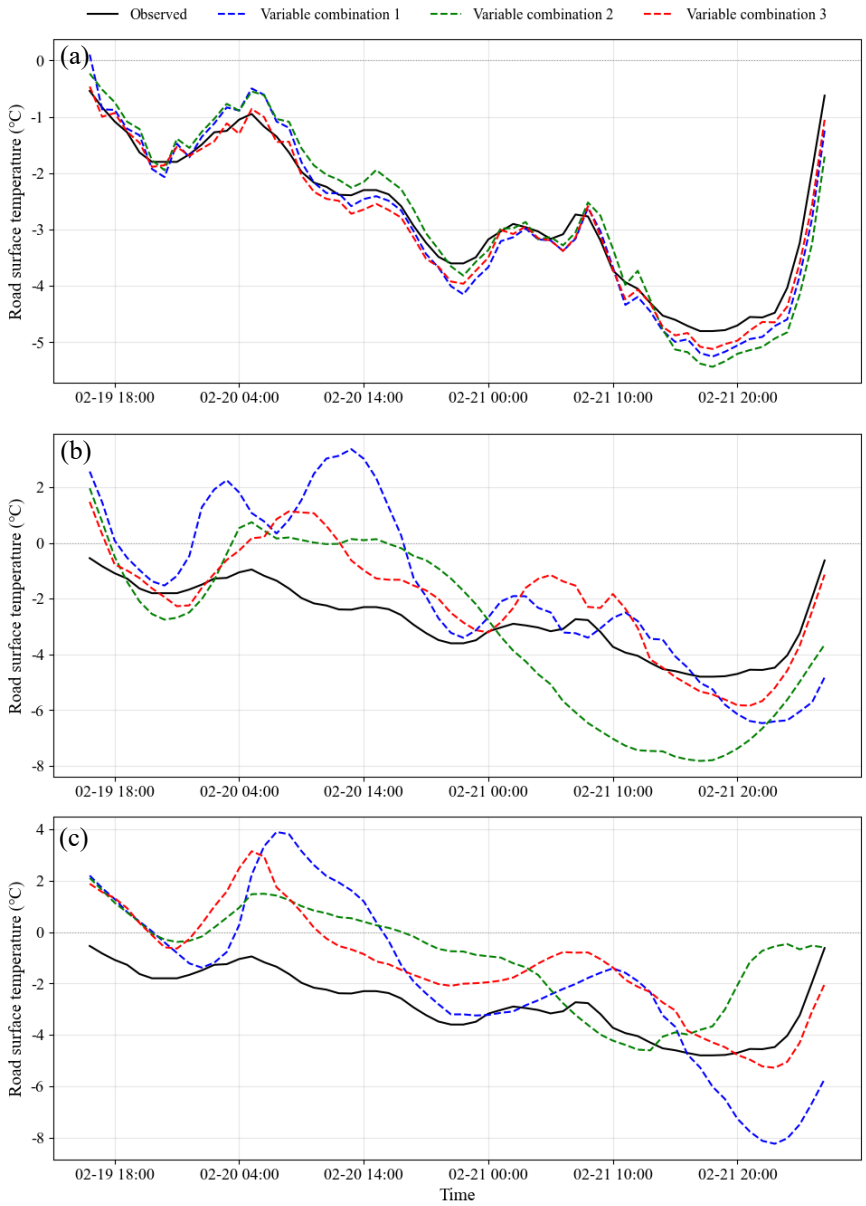
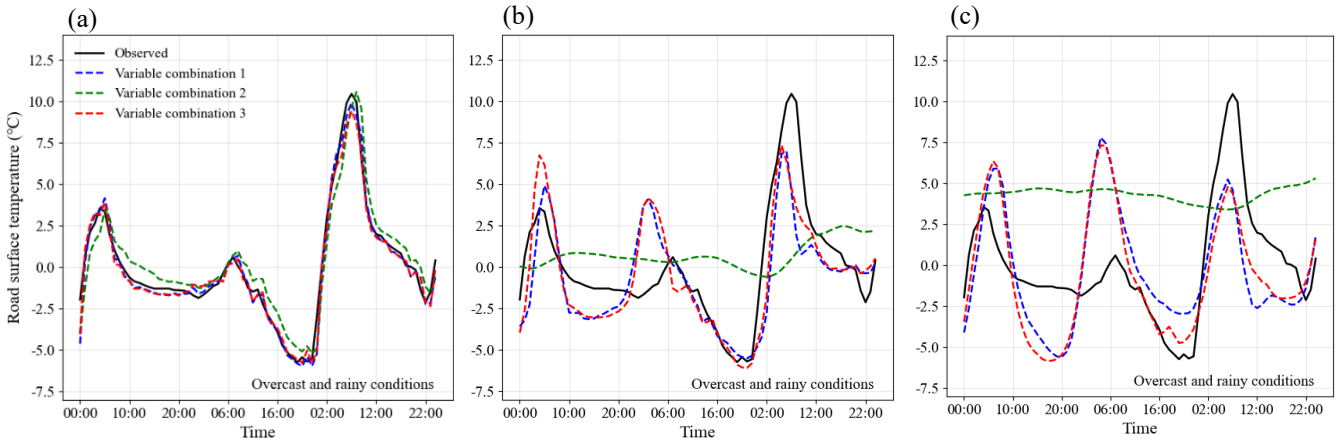


Figure 11. Winter RST prediction of ILES model with three input variable combinations in subzero low temperature period across 1-hour (a), 3-hour (b), and 6-hour (c) forecasting intervals. This period is the longest continuous section of the test sample with RST < 0 °C.



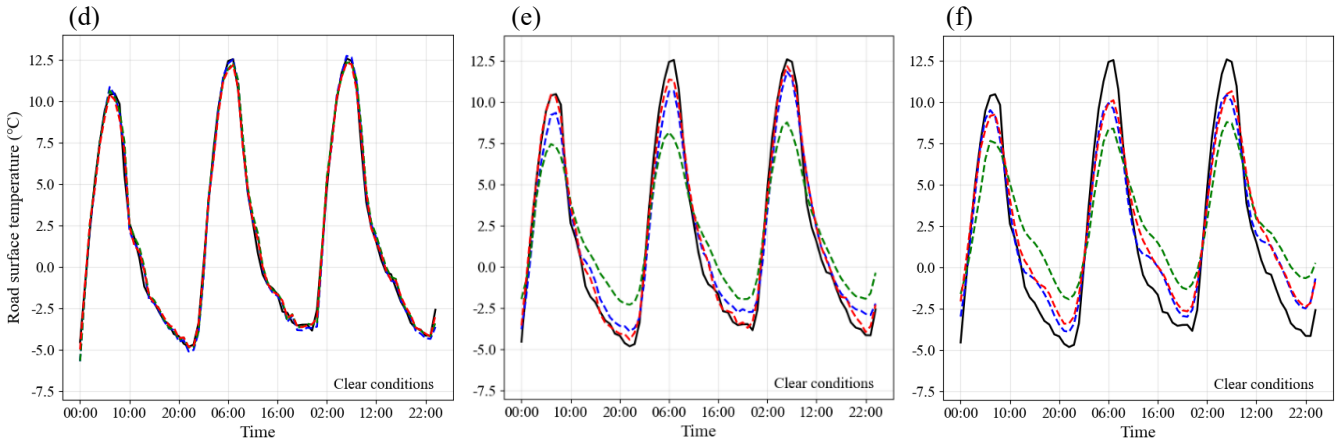


Figure 12. Winter RST prediction of ILES model with three input variables in overcast and rainy and clear synoptic conditions across 1-hour (a, d), 3-hour (b, e), and 6-hour (c, f) forecasting intervals.

Methodological Work Stream 4: Leakage-free stacking with temporally aligned cross-validation

The original manuscript described a stacking ensemble but did not explicitly address the data leakage problem inherent in naive stacking implementations, nor did it justify the cross-validation fold structure in relation to the temporal structure of the training data. The revised manuscript now explicitly formalizes the leakage-free out-of-fold (OOF) meta-learner training procedure (Section 2.3, Figure 4): base learner predictions for the meta-learner training matrix are generated exclusively on validation folds that the corresponding base learner did not observe during training. The 3-fold structure is aligned to complete winter seasons (leave-one-season-out), reflecting the real-world deployment scenario of predicting an unseen winter from prior observations. This temporal alignment ensures that the cross-validation scheme is not merely statistically sound but also methodologically representative of the intended operational use case.

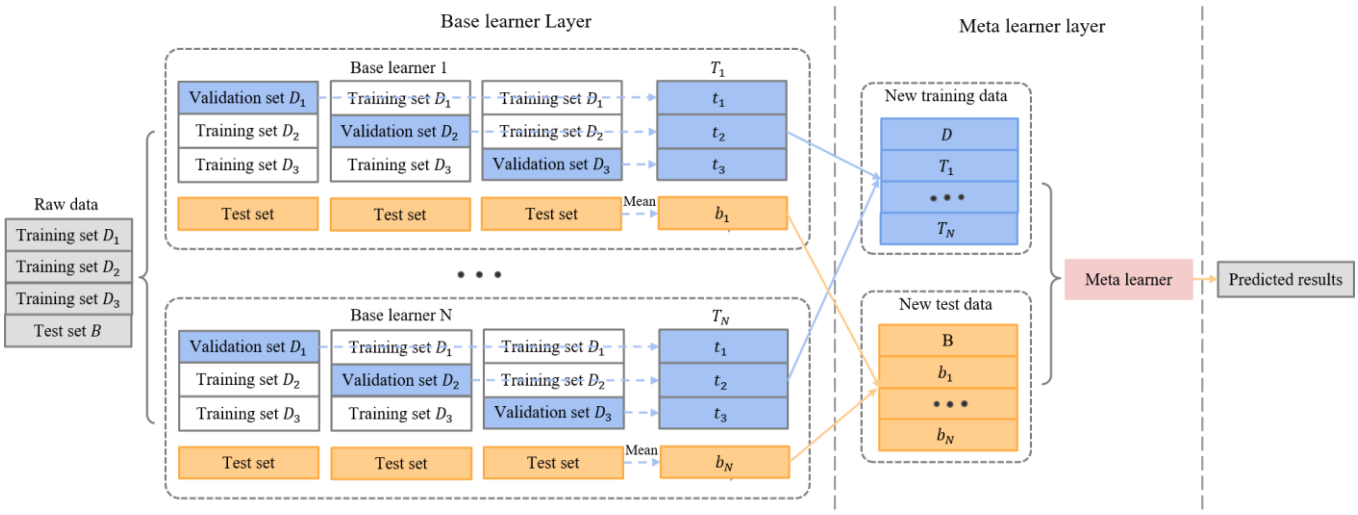


Figure 4. Architecture of the Stacking Ensemble based on Out-of-Fold Cross-Validation.

2. *"Applying established ML components to a new application domain can be valuable, but when novelty is incremental the contribution should be justified primarily by the domain-specific gap it addresses and by operationally meaningful evaluation and discussion. At present, the manuscript focuses on improved error metrics, but it is not yet explicit what unmet need in winter road weather operations the approach resolves. Please clarify whether the main contribution is application-driven, and if so, strengthen the Introduction/related work by explicitly synthesizing prior RST/icing-related prediction approaches already discussed and by stating a concrete gap and objective that this work fills."*

Response:

We thank the reviewer for this focused and constructive comment. We agree that the original manuscript did not sufficiently articulate what prior RST prediction approaches had collectively left unresolved, and that the evaluation was framed primarily

around error metric improvement without explicitly connecting those improvements to concrete unmet needs in winter RST prediction research. The revised manuscript addresses this on two levels: in the Introduction and related work, where we now synthesize prior RST-specific approaches and state clearly what this work aims to resolve; and in the body of the manuscript, where a dedicated model scheme design section explicitly maps each analytical perspective to the specific research question it addresses.

Revision 1: Strengthened Introduction and related work synthesis

The original Introduction reviewed relevant literature in broadly themed paragraphs without synthesizing what prior RST-specific approaches had collectively achieved and, critically, what aspects they had not yet satisfactorily resolved. The revised Introduction restructures this discussion around two persistent structural limitations in the RST prediction literature, which are now stated explicitly:

The first is the parameter identifiability challenge. Physics-based road surface models explicitly solve the surface energy balance (SEB) equation coupled to the one-dimensional heat conduction equation in the pavement substrate. While physically interpretable, their accuracy is critically sensitive to parameters — convective heat transfer coefficient, pavement thermal conductivity, volumetric heat capacity, surface emissivity — that are rarely available at operational road weather stations (Qin and Hiller, 2013; Athukorallage et al., 2023). This well-documented limitation motivates a data-driven alternative that does not require thermal parameter calibration.

The second is the temporal dependency challenge. RST is a strongly autocorrelated variable whose evolution reflects the integrated effect of meteorological forcing over the preceding several hours, encoding multi-hour thermal inertia through the pavement heat storage flux Q_G. Statistical and empirical models impose linearity constraints on an inherently nonlinear meteorology–RST relationship (Yin et al., 2019; Kršmanc et al., 2013). Ensemble tree methods and standard ML approaches address the nonlinearity limitation but treat each input sample independently, and therefore cannot exploit the multi-hour temporal autocorrelation structure of RST time series (Yang et al., 2020; Hatamzad et al., 2022). Recurrent architectures such as LSTM and BiLSTM address this more directly, yet individual architectures tend to emphasise either local pattern recurrence or long-range temporal dependencies — but not both simultaneously. The systematic integration of architecturally distinct LSTM variants through principled meta-learning has received comparatively limited attention for RST prediction, and the consistency of any associated predictive gains across multiple forecasting horizons has not been thoroughly demonstrated in the existing literature.

Beyond these two primary limitations, the revised Introduction identifies two further aspects that prior RST prediction studies have not yet adequately addressed:

The third concerns input design strategy at instrumented sites. Operational road weather stations rarely measure the primary radiative forcing variables prescribed by SEB theory. Prior studies have either omitted such variables entirely or incorporated reanalysis-derived surrogates, yet a direct empirical comparison between reanalysis augmentation and physics-motivated feature engineering constructed directly from existing station observations — in terms of both predictive utility and operational practicality — has not been conducted for winter RST forecasting. This leaves an important practical question unresolved for researchers and practitioners working with instrumented sites that have dense local observational records.

The fourth concerns model interpretability under stratified conditions. SHAP-based attribution has been applied to data-driven RST models in prior work, but typically under global averaging conditions without stratification by temperature regime, or hour of day. Such stratified analysis is needed to evaluate whether learned model behaviour is qualitatively consistent with the dominant meteorological drivers identified by SEB theory, and to identify conditions under which the model may be less reliable — a dimension of interpretability that has not been systematically reported in the existing RST prediction literature.

Revision 2: Explicit motivation-to-analysis mapping in the manuscript body

In response to the reviewer's observation that the manuscript focuses on improved error metrics without articulating what the approach concretely resolves, we have added a dedicated Experimental design (Section 3.5,) that explicitly maps each of the four analytical perspectives to the research question it addresses, rather than presenting them as a sequence of parallel experiments:

The first analytical perspective (Section 4.1) examines whether a stacking ensemble of architecturally complementary LSTM variants achieves consistent and durable predictive gains over ten models spanning naive baselines, traditional machine learning,

standard deep learning, and the two proposed base learners, across 1-, 3-, and 6-hour forecasting horizons — directly addressing the question of whether principled ensemble integration of heterogeneous LSTM architectures provides a more robust solution to the temporal dependency challenge than individual architectures. Uncertainty quantification provided by the BRR meta-learner is additionally assessed through a reliability diagram on the held-out test set.

The second analytical perspective (Section 4.2) examines whether physics-motivated feature engineering — incorporating the surface–air temperature difference T_{grad} and multi-scale RST tendency features constructed from existing station observations — provides a more effective and operationally practical input strategy than ERA5-Land reanalysis augmentation, directly addressing the input design question for instrumented sites.

The third analytical perspective (Section 4.3) examines whether the learned feature importance structure of ILES, as quantified by SHAP analysis stratified by temperature regime and diurnal cycle, is qualitatively consistent with the dominant meteorological drivers identified by SEB theory — directly addressing the interpretability question identified in the related work review.

The fourth analytical perspective (Section 4.4) examines whether the ILES framework maintains its predictive superiority at two independent stations with distinct geographical and thermal boundary conditions, addressing the practical question of whether the proposed methodology transfers beyond the primary training site.

This structure ensures that every result reported in the manuscript is explicitly connected to a specific aspect of prior work that remained insufficiently resolved, rather than functioning solely as an error metric comparison.

On the nature of the contribution

To be precise in response to the reviewer's question about whether the contribution is primarily application-driven: the contribution is methodologically driven with domain-specific motivation. The paper does not claim to have invented fundamentally new ML components. Rather, it demonstrates that: (i) the specific combination of KNN-LSTM and BiLSTM-MHA within a leakage-free stacking ensemble addresses a documented aspect of the temporal dependency challenge that individual architectures have not resolved; (ii) the controlled comparison of input strategies addresses a practical question about how to embed SEB-relevant information into data-driven models without requiring additional sensors or reanalysis access; and (iii) stratified SHAP analysis provides a more rigorous basis for evaluating physical plausibility of learned representations than has been reported in prior RST studies. These contributions are domain-specific in their motivation and evaluation, and methodological in their form — a distinction that is now clearly articulated in the revised manuscript.

3. *"The Introduction includes extensive background (e.g., general LSTM and ensemble learning) but the problem formulation is not defined early enough, so the narrative does not yet flow cleanly as "Background → Research gap → Objective → Contribution." Please consider restructuring so that the paper quickly specifies what is predicted (RST at 1/3/6-hour horizons at hourly resolution), why it is needed (specific winter road maintenance/icing-warning use-cases), what remains unresolved in prior work, and what this study contributes (novelty and value)."*

Response:

We thank the reviewer for this precise structural recommendation. We agree that the original Introduction's narrative did not flow cleanly from background to problem formulation, and that generic LSTM and ensemble learning background occupied substantial space before the RST-specific research problem was clearly defined. The revised Introduction has been completely reorganized to follow the logical structure the reviewer describes: Problem motivation → Prior approaches and their limitations → What remains unresolved → Research objectives and contributions. We describe the restructuring below.

Paragraph 1 (Lines 38–43): Problem motivation and operational context

The revised Introduction opens immediately with RST as the primary indicator for pavement icing events, connecting sub-zero RST directly to reduced friction coefficients, elevated accident risk, and the need for timely anti-icing and de-icing interventions. The specific prediction task — RST at 1-, 3-, and 6-hour horizons at hourly resolution — is introduced early in the Introduction rather than deferred to the Methods section, ensuring the reader immediately understands what is being predicted and why it matters for winter road operations.

Paragraphs 2–3 (Lines 44–76): Physical basis and two fundamental challenges

Rather than opening with general LSTM history, the revised Introduction introduces the surface energy balance (SEB) framework as the physical basis governing RST evolution. This serves two purposes: it grounds subsequent feature engineering choices in domain knowledge, and it immediately motivates the two persistent structural challenges that organize the literature review — the parameter identifiability challenge (physics-based models require thermal parameters rarely available at operational stations) and the temporal dependency challenge (RST exhibits multi-hour autocorrelation that conventional ML approaches cannot exploit). These two challenges now serve as the organizing framework for everything that follows, rather than emerging late in the narrative.

Paragraphs 4–5 (Lines 77–104): Review of prior data-driven approaches and their limitations

The revised literature review is structured around the two challenges identified above, rather than around generic ML taxonomy. Statistical and empirical models are reviewed in terms of their linearity constraints on the meteorology–RST relationship; ensemble tree methods and standard ML approaches are reviewed in terms of their inability to exploit temporal autocorrelation structure; and recurrent deep learning architectures are reviewed in terms of the tension between local pattern recurrence and long-range temporal dependency capture. This organization ensures that each category of prior work is evaluated against the same criteria, and that the limitations identified are directly traceable to the research questions the paper addresses. The original Introduction's extended treatment of general LSTM history (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) and generic stacking background (Wolpert, 1992; Ledezma et al., 2010) that were not connected to the RST prediction problem have been substantially compressed or removed from the Introduction.

Paragraph 6 (Lines 105–119): What prior work has not yet adequately resolved

Following the literature review, the revised Introduction now includes an explicit synthesis paragraph that identifies what prior RST prediction approaches have collectively left insufficiently resolved. Four specific aspects are identified: (i) the systematic integration of architecturally complementary LSTM variants through principled meta-learning for RST prediction; (ii) an empirical comparison between reanalysis augmentation and physics-motivated feature engineering as input strategies at instrumented sites; (iii) the stratified SHAP interpretability analysis across temperature regimes and diurnal cycles; and (iv) multi-site generalizability evaluation at stations with distinct geographical and thermal boundary conditions. Crucially, these are framed not as abstract methodological gaps but as concrete questions that arise directly from the operational context and the limitations of prior approaches reviewed above.

Paragraph 8–9 (Lines 120–136): Research objectives and contributions

The revised Introduction closes with a clearly structured objectives paragraph that maps directly onto the four analytical perspectives of the study. Each objective is stated as a testable research question rather than a vague methodological claim, and the ILES framework is introduced as the proposed approach to addressing these questions collectively. The abstract description of the contribution in the original manuscript ("this paper proposes a novel forecasting framework") has been replaced with specific statements of what the framework does, what it is evaluated against, and what constitutes a meaningful result for each objective.

The revised Introduction also includes Table 1, summarises recent LSTM-based approaches for RST prediction, including the model architectures employed, the input features considered, and representative predictive accuracy reported in each study. s. This table serves two purposes in relation to the reviewer's comment: it demonstrates that the literature review is grounded in RST-specific prior work rather than generic ML background, and it makes the positioning of the ILES framework relative to prior approaches immediately legible to the reader without requiring them to reconstruct it from narrative prose alone.

Table 1. Summary of LSTM-based approaches for road surface temperature prediction. Abbreviations: AT, air temperature; SR, solar radiation; RH, relative humidity; P, precipitation; WS, wind speed; WD, wind direction; AP, atmospheric pressure; Depth, measurement depth below pavement surface; Time, time-of-day index.

Reference	Model	Feature	Interval	Characteristic	Evaluation
Tabrizi et al. (2021)	CNN-LSTM	RST, AT, SR	1h	CNN-LSTM hybrid architecture for multi-horizon forecasting	MAE=1.05–3.43 °C and R^2=0.98–0.80 across 1- to 6-hour horizons
Milad et al. (2021a)	Bi-LSTM	AT, Depth, Time	1h	Bidirectional LSTM with enhanced feature extraction	MAE = 1.332 °C; R^2 = 0.956

				across depth and time dimensions	
Bai et al. (2022)	Att-BiLSTM	RST, AT, P, WS, RH	1h	Attention-augmented BiLSTM with sliding window optimisation for micro-scale prediction	MAE = 0.330 °C; 93.4% of predictions within ±1 °C
Dai et al. (2023)	GRU, LSTM	RST, AT, P, WS, RH	1h	GRU-LSTM ensemble exploiting RST periodicity and meteorological lag effects	MAE of 0.345, 0.833, and 1.743 °C at 1-, 3-, and 6-hour horizons
Zhang et al. (2024)	RF-LSTM	RST, AT, AP, WS, RH, WD	10min	RF-based feature selection integrated with LSTM for short-term prediction	MAE = 0.048 °C; 99.1% within ±0.5 °C
Our paper	ILES	RST, AT, RH, P, WS	1h	Stacking ensemble of KNN-LSTM and BiLSTM-MHA; physics-motivated feature engineering; probabilistic forecasting via BRR	MAE of 0.373, 1.268, and 2.108 °C at 1-, 3-, and 6-hour horizons; R^2 = 0.993–0.826

4. *"Section 3.3.3 states that all variables were standardized using Z-score normalization, but it is unclear whether the target RST was standardized, whether predictions were inverse-transformed before evaluation, and therefore whether the reported MAE/MSE/MAPE values are in ℃ or in standardized units. This matters because the paper reports very small errors (e.g., MAE = 0.074) while figures label MAE in ℃. Please clarify the evaluation scale, and also clarify how μ and σ were computed (training only vs. full dataset). If normalization parameters are computed using the full dataset (including the test winter), that introduces information leakage; please confirm normalization is fitted on training only and applied to validation/test."*

Response:

We thank the reviewer for identifying this critical ambiguity. In the original manuscript, the normalization procedure was insufficiently described, creating legitimate concern about evaluation scale and potential information leakage. The revised manuscript addresses all three sub-issues explicitly.

Issue 1 Evaluation scale: The original manuscript's reported MAE of 0.074 was indeed in standardized units, not in degrees Celsius, which created a misleading impression of very small absolute errors. This has been corrected throughout. In the revised manuscript, all reported MAE and RMSE values are in degrees Celsius following inverse transformation, as now stated explicitly in Section 3.3.2:

"Model outputs were inverse-transformed to degrees Celsius prior to evaluation. Accordingly, all reported MAE and RMSE values are in °C."

Specifically, the 1-hour MAE reported in the revised Table 5 is 0.373°C for ILES, which is the physically meaningful, inverse-transformed value. The discrepancy between the original manuscript's "MAE=0.074" and the revised "MAE = 0.373°C" reflects this correction: the former was in normalized units whereas the latter is in °C as labeled.

Issue 2 Normalization parameters computed from training set only: The revised manuscript now explicitly confirms that normalization parameters are estimated exclusively from the training set and applied to validation and test data without refitting. Section 3.3.2 now reads:

"Prior to model training, all variables were standardized using Z-score normalization to accelerate convergence and eliminate the influence of differing measurement scales. Normalization parameters were estimated exclusively from the training set to prevent information leakage:

$$X_{scaled} = \frac{X_{train} - \mu_{train}}{\sigma_{train}} , \quad (27)$$

where μ_{train} and σ_{train} are the mean and standard deviation computed from the training set only."

The test set was normalized using the training-set parameters (μ_{train}, σ_{train}) without any refitting, which ensures that no information from the test winter (December 2023 to February 2024) enters the normalization procedure. This is the standard

practice for preventing information leakage in time-series model evaluation (Wolpert, 1992; Breiman, 1996).

Issue RST target variable: All input features and the RST target variable were standardized using the same Z-score procedure with training-set parameters. Predictions are inverse-transformed to °C before computing all evaluation metrics. This applies consistently across all three forecasting horizons (1-, 3-, 6-hours), all evaluation metrics (MAE, RMSE, sMAPE, R^2), and all experimental configurations.

Additionally, we note that the revised manuscript has replaced the original MAPE metric with sMAPE (symmetric MAPE, Eq. 24) throughout, which partially addresses the numerical instability associated with near-zero denominators in winter RST datasets, as raised in Reviewer 2's comment. The sMAPE formula 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 $\varepsilon = 10^{-8}$ prevents division by zero when both observed and predicted values simultaneously approach zero.

5. *"Bayesian ridge regression is presented as a key meta-learner and the manuscript states it provides a probabilistic framework that can model uncertainty, but the Results section reports point-estimate metrics only and does not describe predictive uncertainty. Please provide the mathematical formulation used for Bayesian ridge regression (including prior assumptions and how the posterior is obtained), and describe precisely how the stacking meta-learner is trained in your time-series setting, since Section 2.3 states that cross-validated base-learner outputs are used for supervised learning. Please also clarify whether the base learners output deterministic point predictions; if so, explain what "uncertainty" refers to in this implementation and either report predictive intervals (and how they would be used operationally) or temper the uncertainty-related claims. Finally, please explain why this meta-learner is substantively different from a simple weighted linear combination in this specific implementation, beyond regularization, given that only two base-model predictions appear to be combined."*

Response:

We thank the reviewer for this detailed and technically precise critique. The original manuscript's treatment of Bayesian Ridge Regression (BRR) was insufficiently rigorous. The revised manuscript provides a complete mathematical formulation, clarifies the nature of the uncertainty quantification, reports empirical coverage of predictive intervals, describes their operational interpretation, and explicitly justifies the substantive differences from simple weighted linear combination.

Mathematical formulation: Section 2.4 now provides the complete BRR specification:

The BRR meta-learner assumes a linear relationship between base learner predictions and observed RST, with Gaussian noise and hierarchical priors on the weight precision:

$$y = Xw + \epsilon, \epsilon \sim \mathcal{N}(0, \alpha^{-1}I) \ , \tag{17}$$

$$w \sim \mathcal{N}(0, \lambda^{-1}I), \alpha \sim \text{Gamma}(a_0, b_0), \lambda \sim \text{Gamma}(c_0, d_0) \ , \tag{18}$$

The posterior distribution over weights is analytically tractable (Bishop, 2006, Pattern Recognition and Machine Learning, Springer):

$$w \mid y, X, \alpha, \lambda \sim \mathcal{N}(\mu_n, \Sigma_n), \Sigma_n = (\lambda I + \alpha X^{\top}X)^{-1}, \mu_n = \alpha \Sigma_n X^{\top} y \ , \tag{19}$$

The precision hyperparameters α and λ are optimized via Type-II Maximum Likelihood (Evidence Maximization), which automatically balances data fit against regularization without requiring manual cross-validation (MacKay, 1992).

Predictive uncertainty and intervals (Section 4.1): Both base learners output deterministic point predictions for each time step. The uncertainty quantification in ILES resides entirely in the BRR meta-learner, which yields a closed-form posterior predictive distribution for each forecast:

$$p(\tilde{y} \mid x, \mathcal{D}) = \mathcal{N}(\tilde{y} \mid \mu_n^{\top}x, \sigma_n^2(x)) \ , \tag{20}$$

where $\sigma_{\text{n}}{}^2(x) = \alpha^{-1} + x^{\top}\Sigma_{\text{n}}x$ decomposes total predictive variance into aleatoric noise α^{-1} (irreducible observational noise) and epistemic uncertainty encoded in the posterior covariance Σ_{n}. Crucially, this uncertainty is input-dependent: $\sigma_{\text{n}}{}^2(x)$ is larger for test inputs x that lie further from the training distribution, reflecting reduced posterior certainty about the appropriate combination of base learner outputs in that region of input space.

Empirical coverage is evaluated on held-out test sets at all three stations and reported in Section 4.4 with accompanying reliability diagrams (Fig. 16). At M9393, the 68%, 90%, and 95% prediction intervals achieve empirical coverages of 88.1%, 96.8%, and 98.1%, with mean interval widths of 1.523°C, 2.519°C, and 3.000°C. At M9474, the corresponding coverages are 88.1%, 95.6%, and 97.5% with mean widths of 0.967°C, 1.600°C, and 1.906°C. At M9448, coverages of 88.1%, 95.6%, and 97.5% are achieved with mean widths of 2.216°C, 3.666°C, and 4.368°C. The notably narrower intervals at M9474 reflect the lower RST variability characteristic of that bridge-mounted station. Critically, the conservative over-coverage pattern is consistent across all three sites, confirming that it is a structural property of the BRR posterior predictive decomposition rather than a site-specific artefact. This systematic over-coverage is operationally preferable for safety-critical road maintenance: the cost of missing a genuine icing event substantially exceeds the cost of a false alarm. The operational use of the predictive interval is explicitly stated in the revised text: when the lower bound of the 95% predictive interval falls below 0°C, maintenance crews may initiate precautionary pre-treatment irrespective of the point forecast.

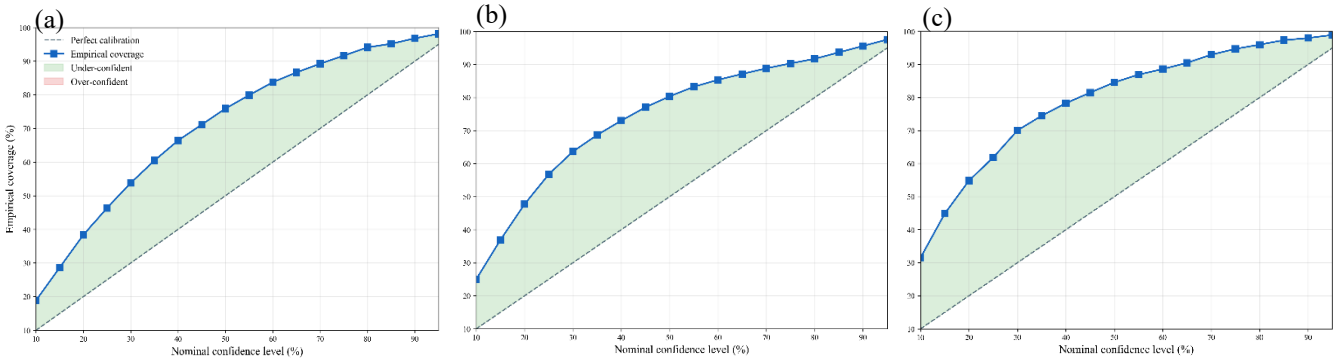


Figure 15. Reliability diagrams of the BRR meta-learner for 1-hour winter RST prediction at (a) M9393, (b) M9474, and (c) M9448.

Substantive differences from simple weighted linear combination: The reviewer correctly notes that with only two base learners, explicit justification is required. Three substantive differences are identified in the revised manuscript (Section 2.4):

First, base learner predictions are correlated due to shared input features. BRR's L2 regularisation via the $(\lambda I + \alpha X^{\mathrm{T}}X)^{-1}$ inversion remains numerically stable even when $X^{\mathrm{T}}X$ is near-singular, a condition that would cause fixed-weight combination to fail or require separate regularisation design choices (Hoerl and Kennard, 1970).

Second, Evidence Maximisation determines the optimal regularisation strength α and λ jointly from data, without requiring manual hyperparameter tuning or a separate held-out validation set. A simple weighted combination requires either fixing the weights a priori or introducing an additional tuning step, both of which consume degrees of freedom that BRR automatically manages.

Third, and most fundamentally, BRR yields an input-conditional posterior predictive distribution rather than a point estimate. The predictive variance $\sigma_{\mathrm{n}}{}^2(x)$ varies across test inputs, reflecting the degree to which the training data inform the combination at that specific operating point. A fixed-weight combination, by contrast, produces a single scalar uncertainty estimate that is constant across all inputs and therefore cannot distinguish between well-supported and extrapolated predictions.

Stacking training procedure in the time-series setting (Section 2.3): The revised manuscript provides a precise description of the leakage-free 3-fold temporal cross-validation procedure (Fig. 4). The three folds correspond to the three complete winter seasons in the training set (2020–2021, 2021–2022, 2022–2023), implementing a leave-one-season-out protocol that mirrors the operational scenario of predicting an unseen winter from prior observations. Fold boundaries are strictly aligned with seasonal transitions and temporal ordering is preserved throughout, ensuring that no future information contaminates the out-of-fold predictions used to train the BRR meta-learner. Both base learners output deterministic point predictions; the stochastic component of ILES arises solely from the BRR posterior predictive distribution at the meta-learning stage.

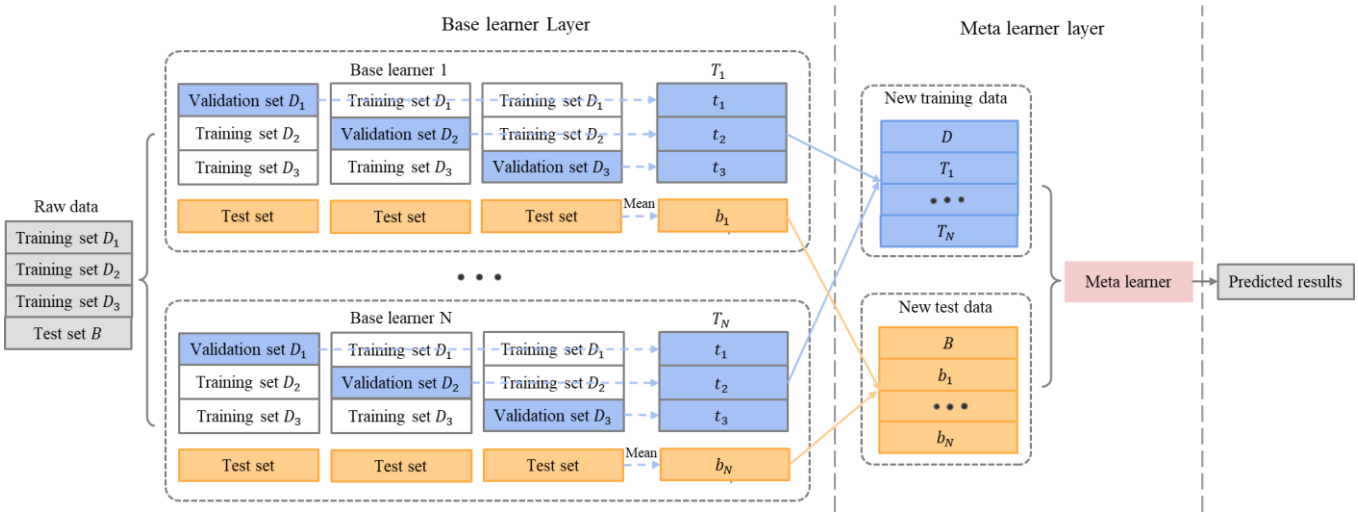


Figure 4. Architecture of the Stacking Ensemble based on Out-of-Fold Cross-Validation.

6. *"Solar radiation is excluded due to missing data, but the discussion under "representative synoptic conditions" emphasizes strong solar radiation effects. If solar radiation is not an explicit input, please clarify how the model is expected to capture radiation-driven diurnal forcing, for example indirectly via lagged RST and the 24-hour input window that encodes the day-night cycle. The discussion should clearly distinguish between external meteorological interpretation and what is actually represented in the model inputs, and the manuscript could briefly note possible proxy features (e.g., time-of-day) as future work while avoiding overinterpretation."*

Response:

We thank the reviewer for this precise and important observation. The original manuscript excluded solar radiation due to missing data but subsequently discussed solar radiation effects under representative synoptic conditions without clearly distinguishing between what is directly represented in the model inputs and what is inferred from meteorological context. The revised manuscript addresses this inconsistency explicitly across three aspects.

Clarification 1: How the model captures radiation-driven diurnal forcing without direct solar radiation input

Solar radiation is not an explicit model input in any of the three input configurations evaluated in this study. However, radiation-driven diurnal forcing is captured indirectly through two mechanisms that are explicitly present in the model inputs, and which the revised manuscript now describes clearly in Sect. 3.2 and Sect. 4.1.

The primary mechanism is the historical RST sequence within the 24-hour sliding input window. RST itself integrates the net effect of all surface energy balance flux terms — including shortwave and longwave radiation, sensible and latent heat fluxes, and ground heat storage — at the pavement surface. The 24-hour window is specifically selected to align with the diurnal solar cycle, enabling the model to learn the day–night thermal rhythm directly from the historical RST trajectory without requiring an explicit radiation measurement. As shown in Fig. 6, RST reaches its mean daily maximum at approximately 14:00 and its minimum during pre-dawn hours, consistent with solar forcing, and this pattern is well-represented in the historical RST inputs available to both base learners.

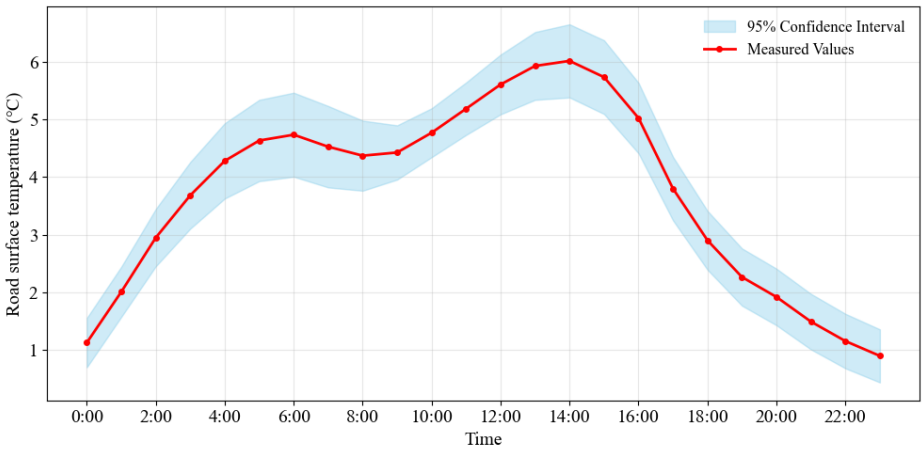


Figure 6. The 24-hour diurnal variation curve of RST. The shaded blue part of the graph is the 95% confidence interval for the RST.

The secondary mechanism is air temperature (AT), which exhibits its own diurnal cycle driven partly by solar heating of the lower atmosphere and is included as an explicit input in all three configurations. While AT does not directly measure shortwave radiation, it carries temporally correlated information about the diurnal radiative environment that the recurrent processing of both KNN-LSTM and BiLSTM-MHA can exploit across the full 24-hour input window.

Importantly, the SHAP analysis in Sect. 4.3 provides empirical support for this indirect encoding. As shown in Fig. 16a, AT maintains the highest feature importance at all hours, with moderately elevated mean absolute SHAP values during both the solar heating window (09:00–16:00) and the nocturnal cooling window (22:00–06:00). This diurnal modulation of AT importance is consistent with the larger surface–air temperature contrast expected during these periods of strong radiative forcing, and confirms that the model has learned to weight AT more heavily precisely when radiation-driven thermal contrast is greatest — even without direct access to radiation measurements.

Clarification 2: Avoiding overinterpretation in the synoptic conditions discussion.

The revised manuscript (Sect. 4.2) now consistently distinguishes between external meteorological interpretation and what is actually represented in the model inputs. The original statement that "strong solar radiation leads to more pronounced temperature fluctuations" — which implied that the model explicitly responds to radiation — has been replaced with formulations that attribute observed RST variability to the diurnal cycle without implying that radiation is a direct model input. For example, the clear-sky performance discussion now reads: " Under clear-sky conditions (Fig. 12d to f), RST exhibits pronounced diurnal oscillations with peak-to-trough amplitudes of approximately 17°C. At the 1-hour horizon, all three configurations achieve comparable accuracy. At 3- and 6-hour, Variable combination 2 shows progressive amplitude underestimation, while Variable combination 3 maintains the closest alignment with the observed diurnal cycle, consistent with the high predictive content of multi-scale tendency features under regular periodic forcing." This formulation describes model behavior in terms of what is represented in the inputs rather than attributing performance to solar radiation directly.

Similarly, the discussion of performance degradation under overcast and rainy conditions has been revised to note that reduced predictability under such conditions is partly attributable to the suppression of the regular diurnal RST pattern that the model exploits as an indirect proxy for radiative forcing — a physically coherent interpretation grounded in the available inputs rather than in variables that are absent from the model.

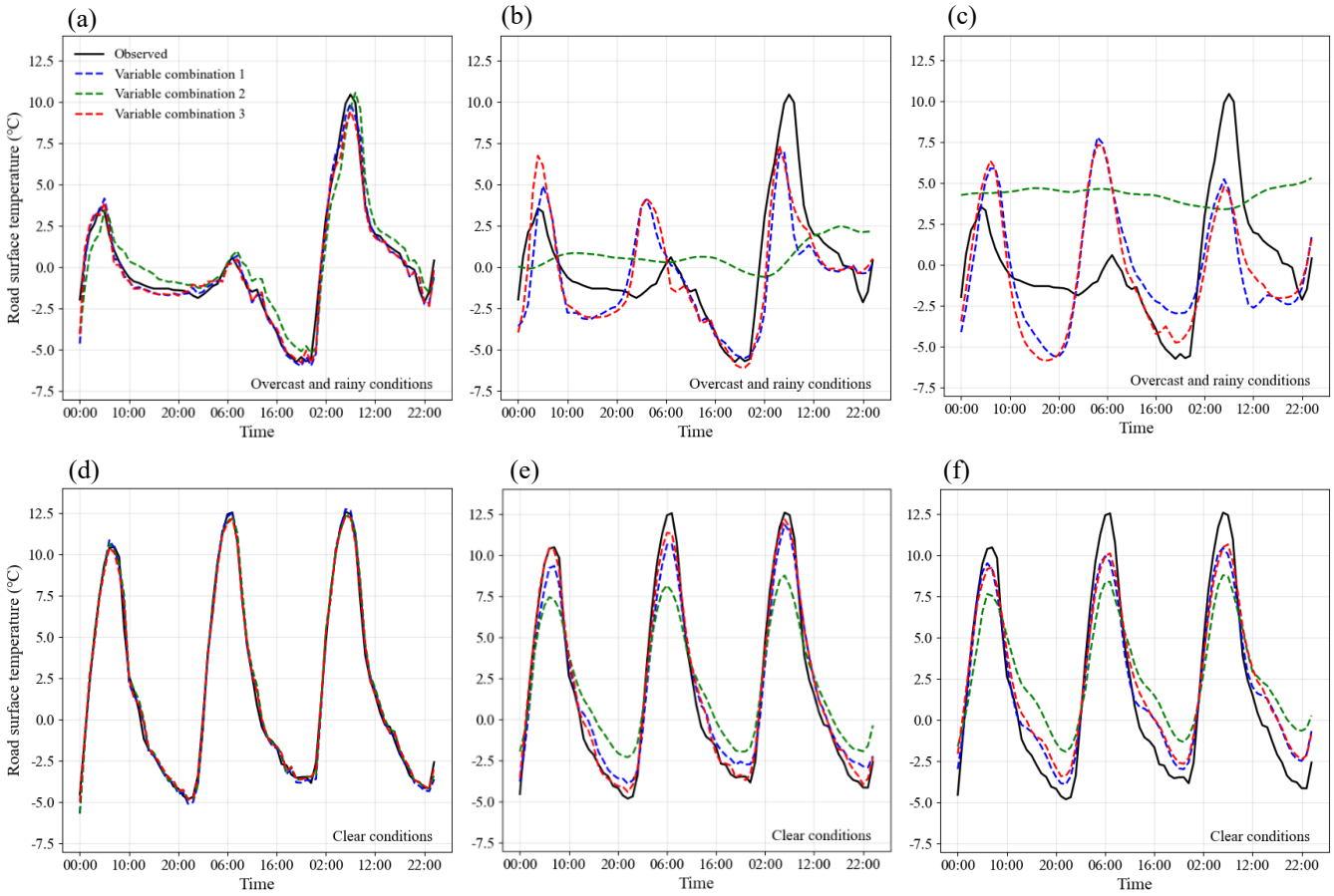


Figure 12. Winter RST prediction of ILES model with three input variables in overcast and rainy and clear synoptic conditions across 1-hour (a, d), 3-hour (b, e), and 6-hour (c, f) forecasting intervals.

Clarification 3: Proxy features as future work.

We agree with the reviewer's suggestion that explicit time-of-day encoding represents a practically straightforward proxy for diurnal radiative forcing that could be incorporated without additional sensor infrastructure. The revised Conclusion (Sect. 5) now notes that future work could investigate the inclusion of cyclical time-of-day encodings — for example, sine and cosine transformations of the hour index — as proxy features for diurnal solar forcing at stations where direct radiation measurements are unavailable. Such features would provide the model with explicit access to the phase of the diurnal cycle independently of the historical RST sequence, and their marginal contribution relative to the implicit encoding already provided by the 24-hour input window could be quantified through controlled ablation experiments analogous to those conducted here for the physics-motivated feature engineering configuration.

Minor Comments

1. *"The Abstract states that the model is trained/validated using "minute-resolution" data, while the Methods describe 5-minute observations that are resampled to hourly intervals for modeling. Please make the Abstract consistent with the actual modeling resolution and forecasting setup."*

Response:

We thank the reviewer for identifying this inconsistency. The reviewer is entirely correct: the primary station M9393 records observations at 5-minute intervals, but all modeling is performed on hourly resampled data. The original Abstract's use of "minute-resolution" was therefore misleading in two respects -it incorrectly implied that the model operates on sub-hourly inputs, and it failed to specify the actual forecasting resolution.

The Abstract has been revised to accurately reflect the data acquisition resolution, the resampling resolution, and the forecasting setup. The relevant sentence in the revised Abstract now reads:

"The framework is trained and evaluated on four consecutive winter seasons of road weather observations from station M9393 in the northwest inland plain of Jiangsu Province, China

This formulation correctly distinguishes between the raw 5-minute acquisition cadence (described in Section 3.1.1) and the hourly modeling resolution, without overstating the temporal granularity of the inputs. The forecasting horizons of 1-, 3-, and 6-hour are explicitly stated in the Abstract, providing readers with an immediate and accurate characterization of the prediction setup. Section 3.1.1 of the revised manuscript now clearly documents both the raw acquisition frequency and the resampling procedure, so there is no longer any ambiguity between the Abstract's characterization and the Methods description.

2. *"The paper states that meteorological factors and RST were resampled at hourly intervals, but the resampling method is not specified (e.g., mean/median/last value; rainfall aggregation). Because this is downsampling, please describe the resampling procedure and any steps taken to avoid introducing artifacts."*

Response:

We thank the reviewer for this important methodological point. The original manuscript described resampling only as "resampled at hourly intervals" without specifying the aggregation operator for each variable type, making the procedure non-reproducible. The revised manuscript now provides variable-specific resampling rules and justifications in Section 3.1.1.

The resampling procedure is as follows. Precipitation was aggregated as the hourly total: summing the twelve 5-minute measurements within each hour correctly recovers the physical hourly depth in millimeters. Using a mean or last-value approach would systematically underestimate hourly accumulation and is inconsistent with the conventions of operational road weather systems (Darghiasi et al., 2023) and the ERA5-Land dataset used in the input configuration comparison. All other variables — air temperature, relative humidity, wind speed, wind direction, visibility, and RST — were computed as hourly means. The mean suppresses transient sub-hourly fluctuations such as brief wind gusts and sensor noise spikes while preserving the underlying thermodynamic state relevant to the surface energy balance process. Using the last value within each hour would introduce arbitrary phase dependence on the position of the observation within the 5-minute cycle and would be more susceptible to single-observation outliers.

To avoid introducing artifacts, hourly means were computed only over valid observations: any 5-minute record flagged as erroneous during quality control was excluded from the within-hour average rather than imputed before averaging. Hours with fewer than 6 valid 5-minute records (corresponding to more than 50% data missing within the hour) were treated as missing at the hourly resolution and subjected to the gap-filling procedure described in Section 3.1.1.

3. *"The manuscript mentions cleaning, outlier removal, and missing-value imputation, but the specific criteria and methods are not described in sufficient detail. Please add reproducible information on thresholds/algorithms and the frequency of these operations."*

Response:

We thank the reviewer for this important reproducibility concern. The original manuscript described data quality control only generically. The revised Section 3.1.1 now provides specific, reproducible criteria for each quality control step:

"Data quality control was applied in two steps. First, physically implausible values were rejected as sensor errors: RST observations exceeding 50°C or falling below −40°C, negative precipitation values, and relative humidity outside the range [0%, 100%] were removed. Second, missing values were imputed according to gap length: gaps not exceeding two consecutive hours were filled by linear interpolation between adjacent valid observations, while longer gaps, occurring primarily during scheduled sensor maintenance, were filled using the climatological mean for the corresponding hour of day computed from the remaining training data."

The full rationale and frequency statistics for each step are as follows:

Step 1 Physical range filtering (outlier removal): The thresholds applied to each variable are grounded in physical plausibility for the study region (temperate monsoon climate, Jiangsu Province):

- RST: valid range [−40°C, 50°C]. The lower bound reflects the absolute minimum RST plausible for the region under any recorded cold event; the upper bound reflects the maximum pavement surface temperature under intense summer solar

radiation (Qin et al., 2022). Values outside this range are attributable to infrared sensor malfunction or transient obstruction.

- Relative humidity: valid range [0%, 100%] (physical definition).
- Precipitation: non-negative (physical definition; negative values indicate sensor drift or calibration error).

The rejection rate was 0% for all variables (see data processing logs), confirming that the station instrumentation was well-maintained and that the quality control step is conservative rather than aggressive.

Step 2 Missing-value imputation: After Step 1, hourly resampling, and filtering of winter months (December, January, February), the resulting hourly winter dataset contained 344 missing records (4.09% of 8,664 hourly samples). These were distributed as follows:

- 26 missing values for V (visibility), 26 for AT (air temperature), 26 for RH (relative humidity) — all part of short gaps (≤2 consecutive hours), imputed by linear interpolation between adjacent valid values;
- 35 missing values for WS (wind speed), 17 for RST (road surface temperature) — part of short gaps (≤2 consecutive hours), imputed by linear interpolation;
- 60 missing values for V, 60 for AT, 60 for RH, 67 for WS, 51 for RST — part of longer gaps (≥2 consecutive hours), primarily during scheduled sensor maintenance, imputed using the climatological mean for the corresponding hour of day computed from the non-missing training data;
- 102 missing values for WD (wind direction) — filled by vector component imputation, as wind direction is a vector quantity unsuitable for linear interpolation or climatological mean imputation.

The choice of linear interpolation for short gaps (≤2 consecutive hours) is standard in road weather meteorology (Nowrin and Kwon, 2022, Cold Regions Science and Technology, 202, 103631) and introduces minimal distortion for gaps of 1–2 hours given the smooth thermal dynamics of RST at hourly resolution. Climatological mean imputation for longer gaps avoids the risk of extrapolating misleading trends over multi-hour periods while preserving the diurnal structure of the hourly mean climatology (Tabrizi et al., 2021, Journal of Hydrology, 603, 126877). Vector component imputation for WD ensures the integrity of wind direction data, as it accounts for the circular nature of this variable.

This complete description of the quality control procedure, including explicit thresholds, imputation algorithms, and frequency statistics, is sufficient for full reproducibility.

4. *"The manuscript states that Spearman correlation is suitable for assessing "non-linear dependencies," but Spearman primarily captures monotonic relationships; please rephrase or justify this statement. In addition, wind direction is a circular variable (0 °= 360 °), so correlation computed directly on degrees can be misleading; even though wind direction is not selected, this limitation should be acknowledged in the feature-selection description."*

Response:

We thank the reviewer for these two precise and well-founded methodological observations. Both points are accepted and have been addressed in the revised manuscript.

Clarification 1: Rephrasing of "non-linear dependencies" in the description of Spearman correlation

The reviewer is correct that Spearman rank correlation captures monotonic relationships rather than non-linear dependencies in general. A non-monotonic non-linear relationship (for example, a U-shaped dependence) would not be detected by Spearman correlation even if the association is strong. The original phrasing was therefore technically imprecise. The relevant sentence in Section 3.2 has been revised to read: "This nonparametric metric ranges from −1 to 1 and is suitable for quantifying monotonic associations between variables, including those that are nonlinear but monotonically ordered." This formulation accurately describes what Spearman correlation detects and avoids overstating its generality as a non-linearity measure.

Clarification 2: Circular nature of wind direction and acknowledgement of the associated limitation

The reviewer correctly identifies that wind direction is a circular variable defined on a [0°, 360°) domain where 0° and 360° are identical. Computing Spearman rank correlation directly on raw degree values imposes a spurious linear ordering on a circular quantity: for example, directions of 350° and 10° are physically separated by only 20° but numerically separated by 340° in the raw representation. This can produce misleading correlation estimates, particularly when observations are concentrated near the 0°/360° boundary.

We acknowledge this limitation explicitly in the revised Section 3.2. The revised text reads: " It is additionally noted that

wind direction is a circular variable for which rank-based correlation computed on raw degree values can be misleading; however, the low correlation magnitude observed here, combined with the absence of directional RST patterns in the sector-based analysis, confirms that the exclusion of wind direction is robust regardless of the association measure applied."

To further support the exclusion of wind direction from the feature set, we conducted a supplementary analysis in which wind direction observations were classified into eight compass sectors (N, NE, E, SE, S, SW, W, NW), each spanning 45° and centred on the standard compass directions, following standard practice in road weather meteorology. The RST distribution within each sector was examined using boxplots (Fig. S1 in the revised manuscript). The analysis shows no systematic pattern in RST median or interquartile range across the eight sectors: median RST values range from approximately 2.00°C to 4.18°C across sectors, overlapping interquartile ranges are observed in all cases, and no sector exhibits a distribution that is statistically distinguishable from the others in terms of central tendency or dispersion. This absence of a directional RST signal is physically consistent with the study site: the Longhai Railway Bridge is situated on flat terrain in the northwest inland plain of Jiangsu Province, where orographic channelling of wind from specific directions is minimal, and the surface energy balance at the pavement surface is not strongly modulated by wind direction per se but rather by wind speed (which is captured as a separate feature). These findings, combined with the low Spearman correlation magnitude reported in Figure 7, confirm that the exclusion of wind direction from the feature set is appropriate and robust to the circularity limitation identified by the reviewer. We also note that wind direction has been similarly excluded in several recent RST prediction studies employing data-driven approaches at comparable station configurations (Dai et al., 2023; Kebede et al., 2024; Zhang et al., 2023), providing further corroboration for this decision.

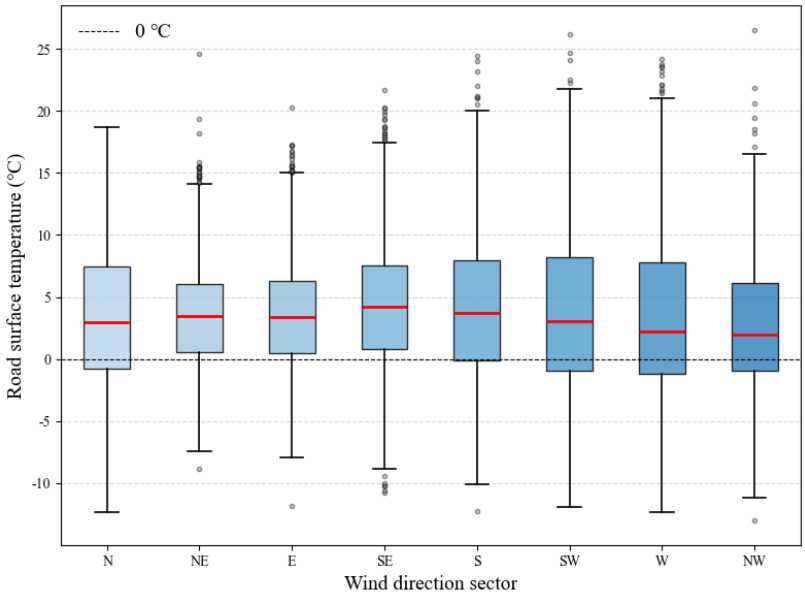


Figure S1. RST Distribution by Wind Direction Sector (8-sector classification, winter seasons 2020-2024).

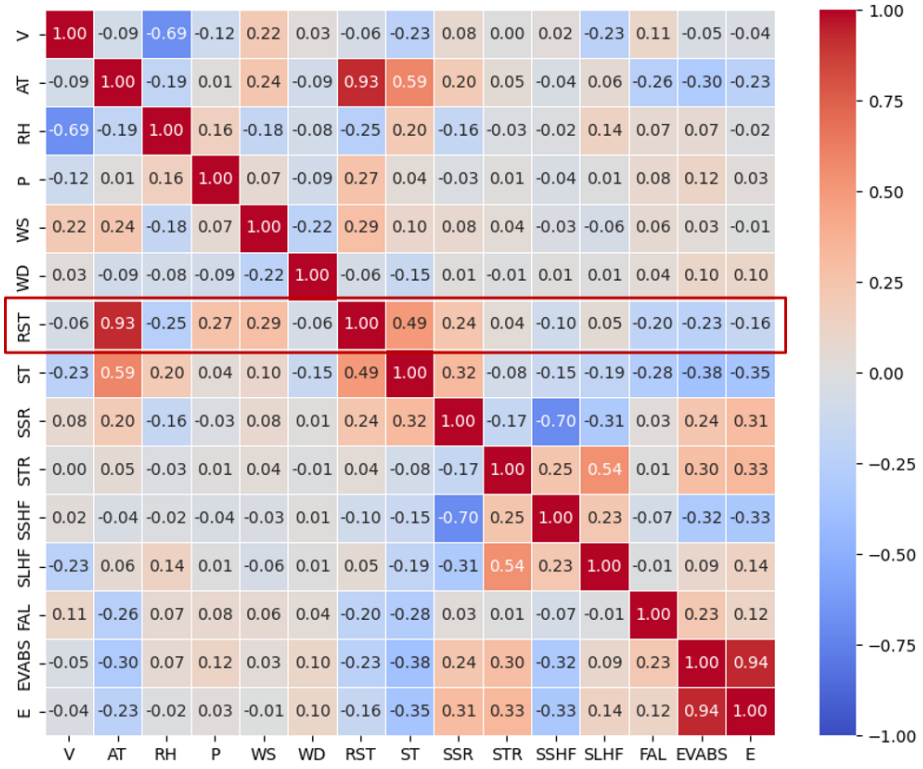


Figure 7. Spearman correlation coefficient plot between RST and various meteorological factors and ERA5_Land physical factors.

5. *"The manuscript evaluates 1-hour, 3-hour, and 6-hour forecasting intervals, but the method section does not clearly explain how the 3-hour and 6-hour forecasts are generated without using future observations. For example, Eq. (3) defines the KNN-LSTM output as the predicted RST at t+1, and the Attention-BiLSTM formulation does not explicitly state how the forecast horizon is handled. Please clarify whether separate direct models are trained for each horizon or whether longer-horizon forecasts are generated recursively, and describe what information is available at time t for the 6-hour forecast. From an information-leakage perspective, future observed values (future RST and future observed meteorological variables) should not be used to generate forecasts beyond time t; if future meteorological inputs are needed, please clarify whether weather-forecast/NWP products are used or what alternative assumption is made. For the BiLSTM-based model, please also clarify that the backward pass operates only within the historical input window up to time t and does not access observations beyond time t, to avoid confusion regarding leakage."*

Response:

We thank the reviewer for raising this critical methodological concern. The original manuscript was indeed unclear on the multi-horizon forecasting strategy, creating legitimate concerns about potential information leakage. The revised manuscript addresses all three sub-issues through a dedicated new section.

Multi-horizon forecasting strategy MIMO: A new Section 3.5 has been added to the revised manuscript to explicitly describe and justify the multi-horizon forecasting strategy. The Multi-Input Multi-Output (MIMO) approach is adopted, in which a single model simultaneously predicts all target horizons (1-, 3-, and 6-hour ahead) within a single inference pass. Formally:

$$(Y_{N+1}, Y_{N+2}, \cdots, Y_{N+k}) = f(X_1, X_2, \cdots, X_N) \ , \quad (21)$$

where $X_1, X_2, \cdots, X_N$ denote the input feature vectors over the 24-hour historical window, $Y_{N+1}, Y_{N+2}, \cdots, Y_{N+k}$ represent the multi-step output targets at horizons $k \in \{1,3,6\}$ hours ahead, and $f(\cdot)$ denotes the predictive model.

The MIMO strategy is chosen over three alternatives (direct, recursive, and direct-recursive hybrid), which are also described in Section 2.5, for two reasons. First, unlike the recursive method, MIMO fundamentally eliminates error accumulation across forecast horizons because predicted values from shorter horizons are never fed back as inputs for longer-horizon predictions. Second, unlike the direct method (which requires training separate independent models for each horizon), MIMO learns shared representations across all horizons within a single model, improving computational efficiency and exploiting the shared

temporal structure across horizons (Ben Taieb et al., 2012).

No future observations are used at inference time: At prediction time t, the model receives only the 24-hour historical input window $\{X_{t-23}, X_{t-22}, \dots, X_t\}$ comprising observed meteorological variables and RST up to and including time t. The model produces predictions for time steps $t+1$, $t+3$, and $t+6$ simultaneously from this single input. No observed values from $t+1, t+2, \dots, t+6$ are used as inputs. No NWP or weather forecast products are used: the 24-hour historical window is sufficient because the MIMO output layer maps directly from historical observations to future targets without autoregressive feedback. This design choice avoids the operational dependency on external forecast products and ensures that the model is deployable at any instrumented road weather station using only historical station records.

BiLSTM backward pass operates only within the historical window: The revised Section 2.2 explicitly clarifies this point:"It should be noted that the backward LSTM pass operates exclusively within the historical input window $[t-T+1, t]$, processing the reversed input sequence without accessing any observations beyond time t. No future meteorological values are used as model inputs."

The backward LSTM processes the reversed sequence $(X_t, X_{t-1}, \dots, X_{t-T+1})$, where all elements are observed at or before time t. The "backward" terminology refers to the direction of processing within the historical window, not to access of future observations. This is the standard formulation of BiLSTM for causal time series forecasting (Schuster and Paliwal, 1997), and the confusion with information leakage is a common one that we now proactively address in the manuscript.

Information available at time t for the 6-hour forecast: The complete information set available at inference time $t$is: observed meteorological variables (AT, RH, P, WS) and RST at hours $t, t-1, \dots, t-23$, comprising 24 hourly observations of 5 variables. No future observations of any variable are available or used. The model predicts RST at $t+1$, $t+3$, and $t+6$purely from these 24×5 historical inputs. This is operationally meaningful: a road weather station operator at time $t$has access to precisely this information and can generate probabilistic RST forecasts at 1-, 3-, and 6-hour lead times without any additional data sources.

6. *"Section 4.3 analyzes two "representative" periods (Feb 8-10, 2024 and Feb 23-25, 2024), but the objective criteria for selecting these periods and their operational representativeness are not clearly described. Please explain how "stable synoptic" and "overcast/rainy" conditions were defined and why these cases were chosen, and clarify how this analysis complements the dedicated sub-zero evaluation in Section 4.2 (i.e., what additional insight it provides for winter road management). In addition, Section 4.3 attributes behavior to strong solar radiation, while solar radiation is excluded from inputs due to missing data; this linkage should be explained carefully (see Major Comment 6)."*

Response:

We thank the reviewer for this multi-part comment, which touches on three distinct aspects of Section 4.3: the selection criteria for the two representative periods, the complementarity of this analysis with the sub-zero evaluation in Section 4.2, and the attribution of model behaviour to solar radiation despite its absence from the model inputs. We address each in turn.

Clarification 1: Objective criteria for period selection and meteorological classification

The original manuscript described the two selected periods as "stable synoptic conditions" and "overcast and rainy conditions" without providing the meteorological criteria used to identify and classify them. The revised manuscript now documents these criteria explicitly in Section 4.2.

The stable clear-sky period (25 to 27 January 2024) was selected on the basis of the following objectively verifiable criteria: mean relative humidity below 50% across the three-day window, zero accumulated precipitation, and mean wind speed below 2 $m \cdot s^{-1}$. These thresholds collectively indicate the absence of cloud cover, precipitation, and strong advective forcing — conditions under which the surface energy balance at the pavement is dominated by shortwave and longwave radiative exchange and the diurnal RST cycle is most pronounced and regular.

The overcast and rainy period (23 to 25 February 2024) was selected on the basis of: continuous precipitation recorded across all three days, mean relative humidity exceeding 85%, and suppressed diurnal RST amplitude (peak-to-trough range below 3°C, compared to approximately 17°C under clear-sky conditions). These criteria identify a meteorological regime in which latent heat fluxes and precipitation-driven thermal damping dominate the surface energy balance, producing RST variations that are more subdued, irregular, and less amenable to prediction from the available observational inputs.

These two periods were identified by screening all three-day windows in the 2024 winter test set against the above criteria and selecting the window that most clearly satisfied each set of conditions. They were chosen to represent the two ends of the winter meteorological spectrum at the study site — periodically dominated versus stochastically driven RST variability — rather than to characterize average winter conditions. Both periods are drawn exclusively from the test set and were not used in any aspect of model training or validation.

Clarification 2: What this analysis contributes beyond the sub-zero evaluation in Section 4.2

The reviewer correctly asks how the synoptic regime analysis complements the sub-zero evaluation. These two analyses address distinct and non-overlapping dimensions of model performance, as clarified in the revised manuscript.

The sub-zero evaluation in Section 4.2 focuses on the accuracy of RST prediction when the target variable falls below the icing threshold (RST < 0°C), which is the operationally critical regime for road safety decisions. Its primary purpose is to assess whether the model's advantage over benchmarks is preserved under the most safety-relevant temperature conditions, regardless of the meteorological forcing regime responsible for those temperatures.

The synoptic regime analysis in Section 4.2, by contrast, examines how model performance varies as a function of the meteorological forcing regime — specifically, the contrast between periodically dominated (clear-sky) and stochastically driven (overcast/rainy) RST variability. This distinction is operationally meaningful for a different reason: it determines under which synoptic conditions the model's predictions can be trusted most reliably, and under which conditions additional caution is warranted in operational deployment. The finding that all models, including ILES, show degraded performance under precipitation-dominated conditions is not captured by the sub-zero evaluation alone, since sub-zero conditions can occur under both clear-sky and overcast regimes. Together, the two analyses provide complementary information: the sub-zero evaluation establishes that the model performs well when it matters most from a safety perspective, while the synoptic regime analysis characterizes the meteorological conditions under which the model's reliability is highest and lowest — information directly relevant to practitioners deciding when to trust automated RST forecasts and when to apply additional verification. This complementarity is now stated explicitly at the end of Section 4.2 of the revised manuscript, where we note that the sub-zero evaluation establishes performance under the most safety-critical temperature regime, while the synoptic regime analysis characterises the meteorological conditions under which model reliability is highest and lowest.

Clarification 3: Attribution of model behaviour to solar radiation despite its absence from model inputs

The original manuscript's Section 4.3 contained the statement that "the presence of stable weather and **strong solar radiation** leads to more pronounced temperature fluctuations," which implied that the model has direct access to solar radiation information. This statement was potentially misleading and has been **removed entirely** from the revised manuscript.

In the revised manuscript, the discussion of clear-sky performance is rephrased to focus on what is physically observed in RST (pronounced diurnal oscillations with peak-to-trough amplitudes of approximately 17°C) and on what the model actually learns from its inputs (the regularity and high predictive content of the historical RST and meteorological sequence under periodic forcing). Specifically, the revised text reads: "Under clear-sky conditions (Fig. 13d to f), RST exhibits pronounced diurnal oscillations with peak-to-trough amplitudes of approximately 17°C... Variable combination 3 maintains the closest alignment with the observed diurnal cycle, consistent with the high predictive content of multi-scale tendency features under regular periodic forcing."

For the degraded performance under overcast and rainy conditions, the revised manuscript explains that precipitation suppresses the regular diurnal RST cycle, which reduces the informativeness of the historical RST sequence as an indirect proxy for the current radiative state of the surface. This explanation is physically coherent and does not require solar radiation to be an explicit model input. No direct causal attribution to solar radiation is made anywhere in the revised manuscript.

It should be noted that ERA5-Land surface net solar radiation (SSR) is included in Variable combination 2, but not in the standard station-only configuration (Variable combination 1) used for primary benchmarking, nor in the physics-motivated configuration (Variable combination 3). Any discussion of solar forcing in the context of model inputs is therefore explicitly restricted to the Variable combination 2 analysis in Section 4.2.

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).

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

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

$$R = H + \mathrm{MultiHead}(H, H, H), \quad (10)$$

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

- 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.

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

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:

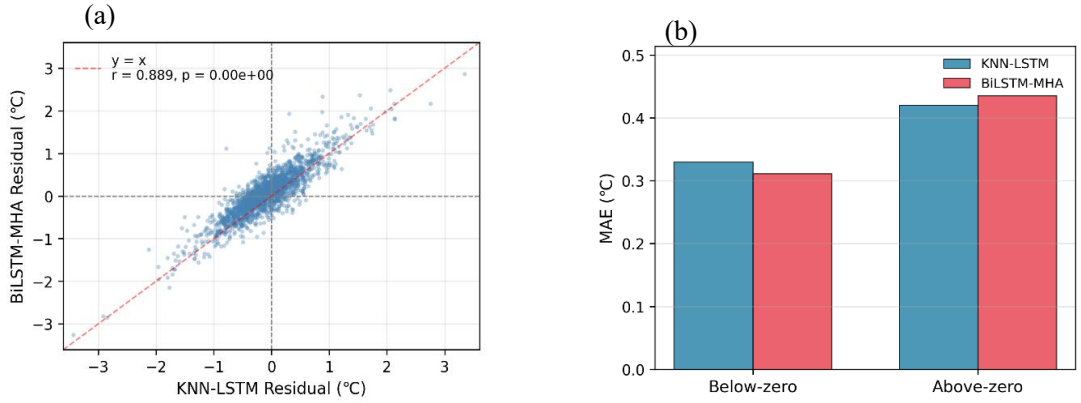
In response to the reviewer's suggestion that a rigorous analysis of residual correlation structure and error decomposition would strengthen the novelty claim, Section 3.4 and Fig.9 now present a four-panel complementarity analysis conducted at the 1-hour forecasting horizon, where the ensemble signal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

Residual correlation (Fig. 9a): The Pearson correlation coefficient between the residual series of KNN-LSTM and BiLSTM-MHA is $\mathrm{r} = 0.889$, 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.



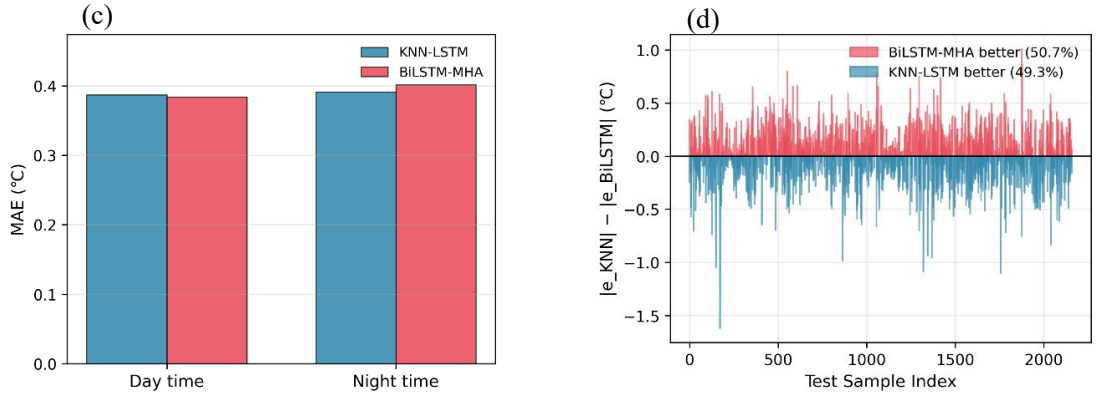


Figure 5. Complementarity analysis of base learners. Where (a) residual correlation analysis, (b) temperature interval error decomposition, (c) day and night time error decomposition, (d) time series advantage distribution.

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.

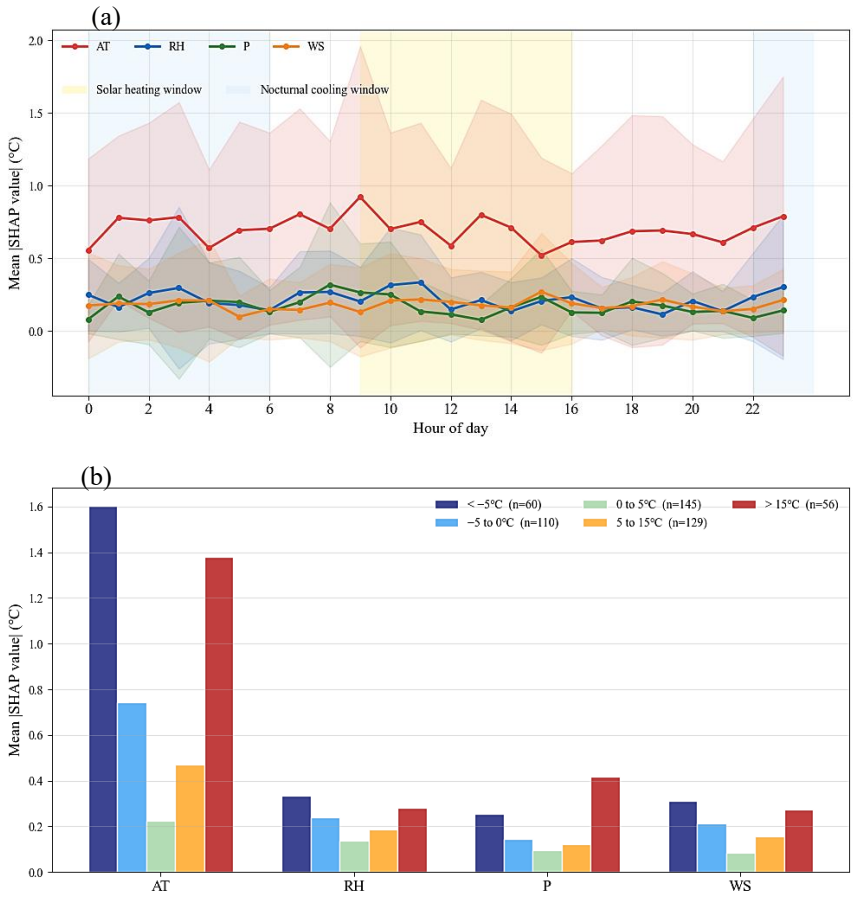


Figure 14. Diurnal variation of mean absolute SHAP values across the 24-hour cycle (a). Shaded regions indicate the solar heating window (09:00-16:00) and nocturnal cooling window (22:00-06:00). Mean absolute SHAP values stratified by RST regime (b). Numbers in parentheses indicate sample counts per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

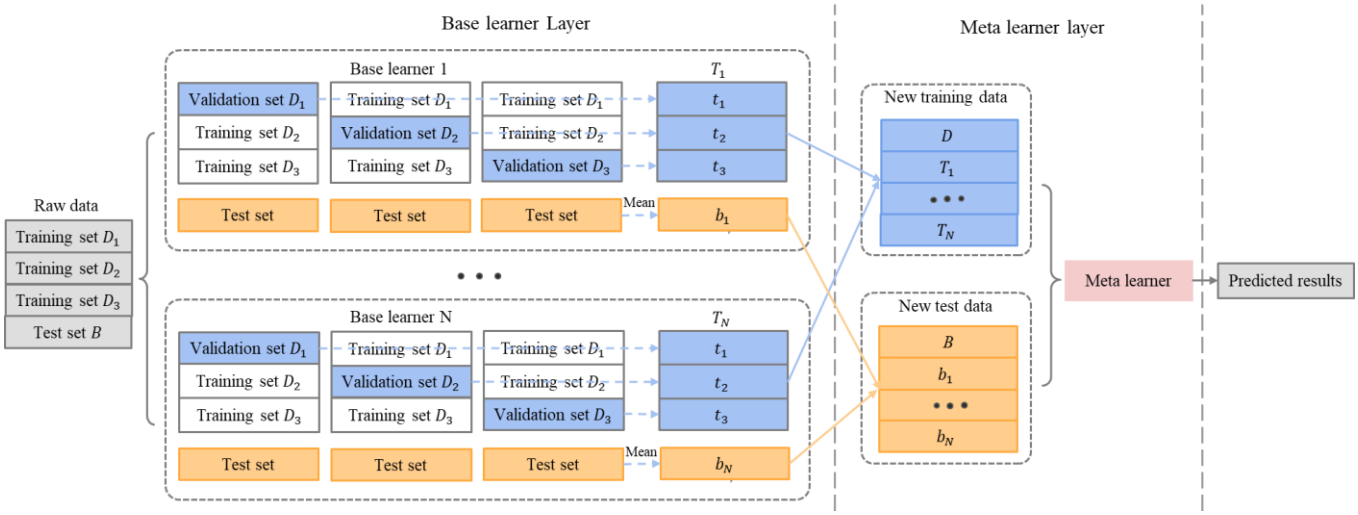


Figure 4. Architecture of the Stacking Ensemble based on Out-of-Fold Cross-Validation.

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.

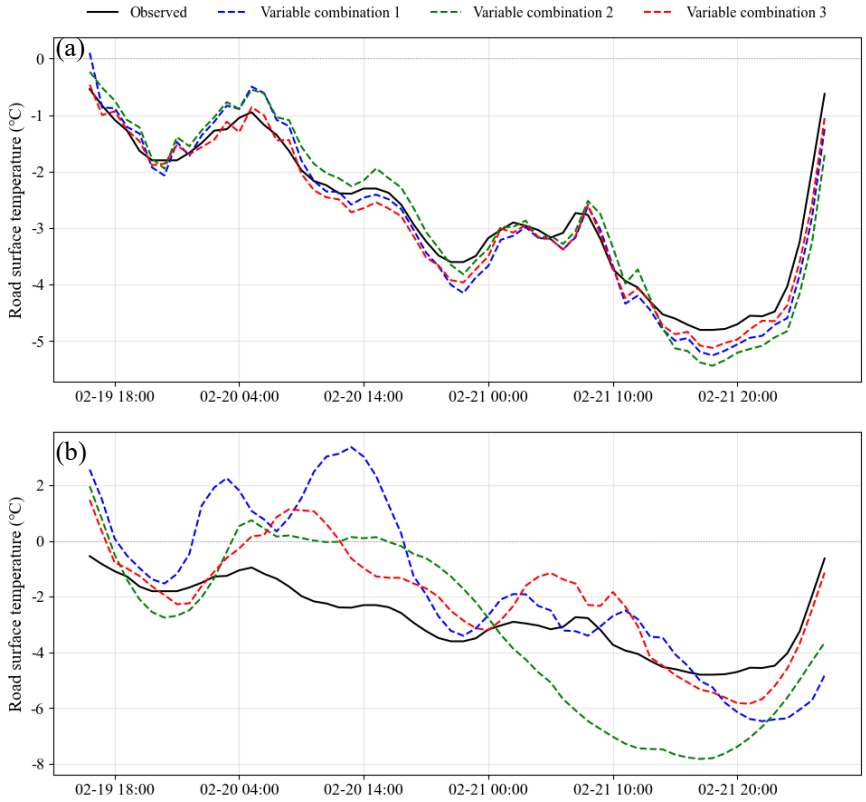
Step 1: Strengthened geoscientific motivation and domain-knowledge-guided feature engineering

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



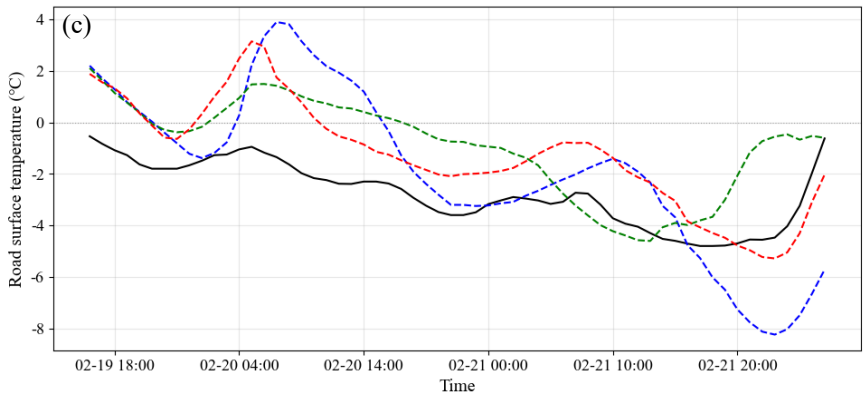


Figure 11. Winter RST prediction of ILES model with three input variable combinations in subzero low temperature period across 1-hour (a), 3-hour (b), and 6-hour (c) forecasting intervals.

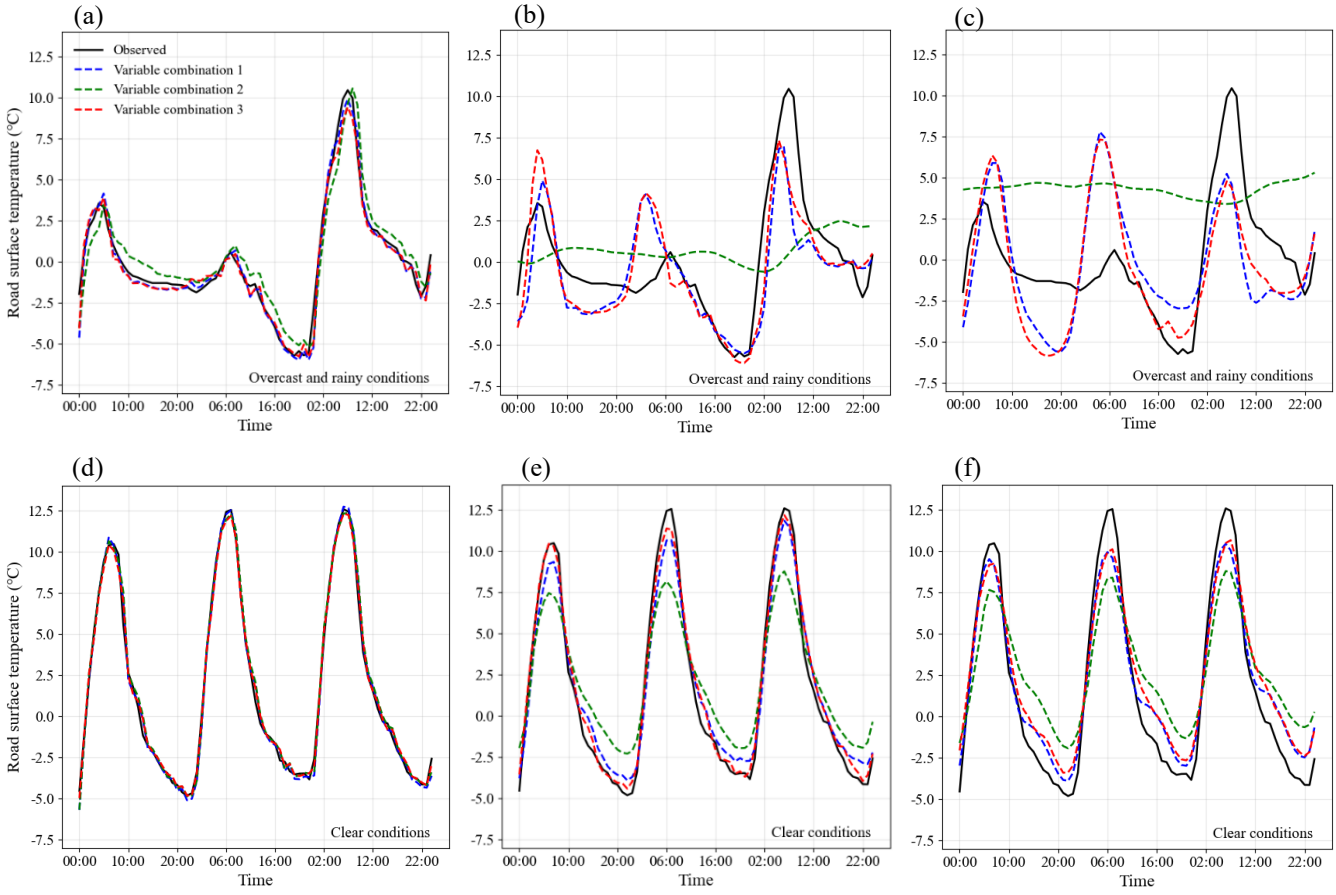


Figure 12. Winter RST prediction of ILES model with three input variables in overcast and rainy and clear synoptic conditions across 1-hour (a, d), 3-hour (b, e), and 6-hour (c, f) forecasting intervals.

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.

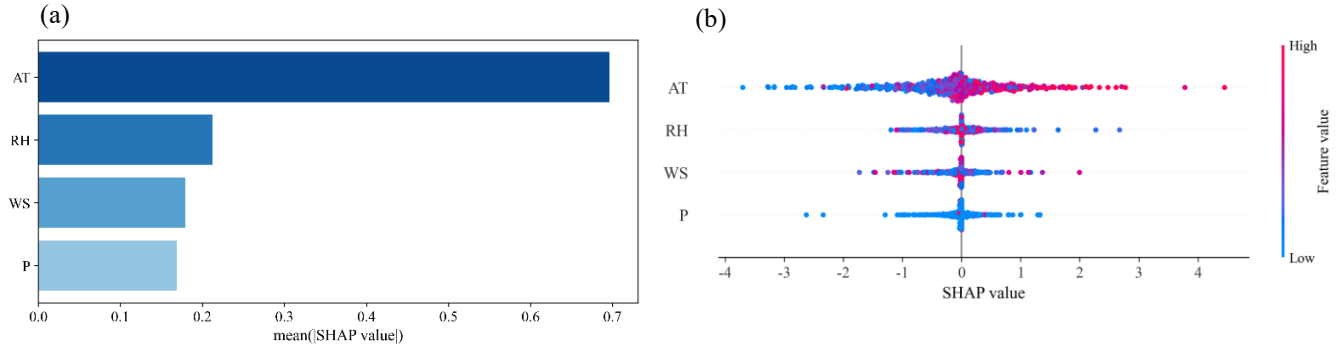
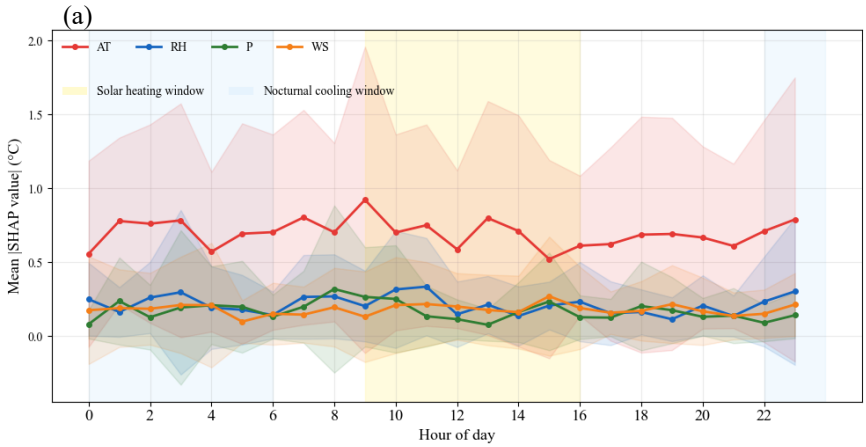


Figure 13. Mean absolute SHAP values (a) and beeswarm plots of SHAP value distributions (b) for the ILES model.

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.



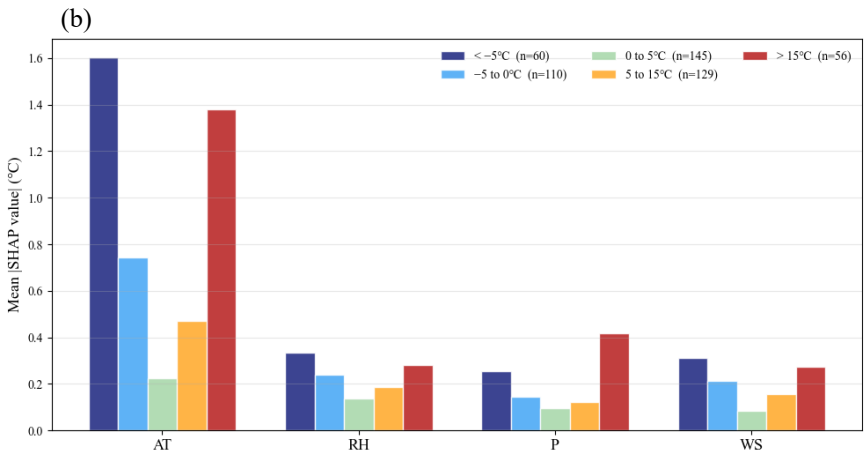


Figure 14. Diurnal variation of mean absolute SHAP values across the 24-hour cycle (a). Shaded regions indicate the solar heating window (09:00-16:00) and nocturnal cooling window (22:00-06:00). Mean absolute SHAP values stratified by RST regime (b). Numbers in parentheses indicate sample counts per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

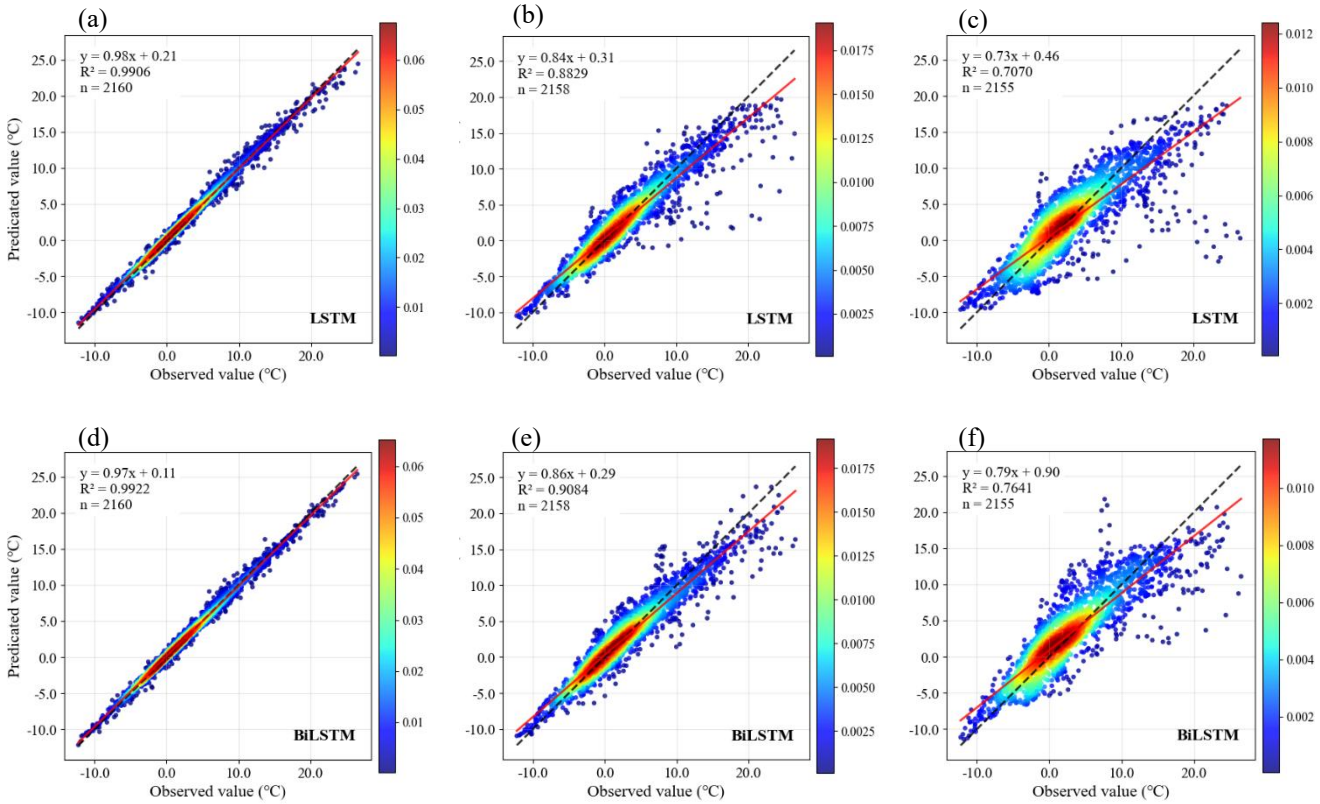
Response:

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.



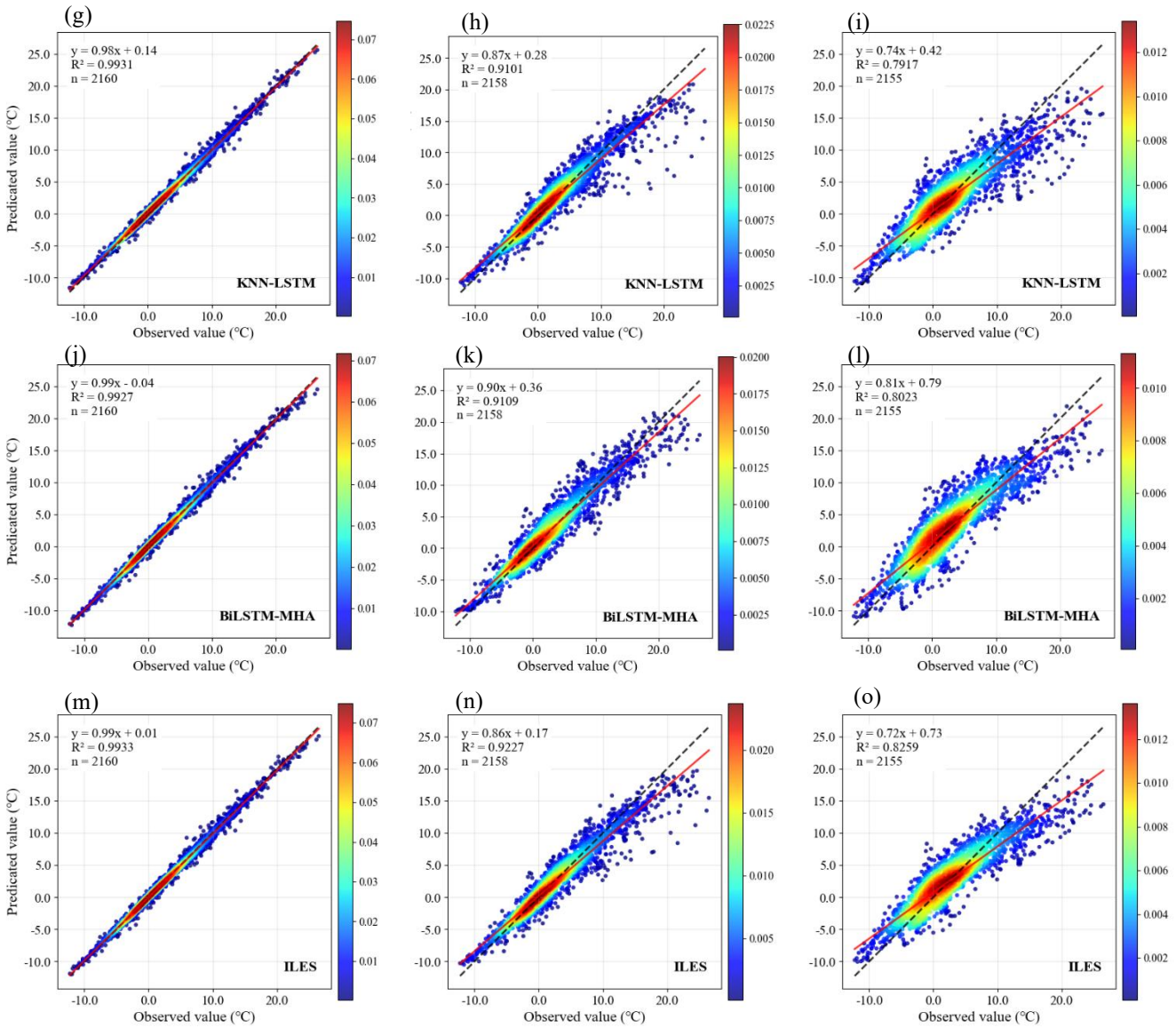


Figure 10. Density scatter plots of predicted versus observed winter RST across 1-hour (a, d, g, j, m), 3-hour (b, e, h, k, n), and 6-hour (c, f, i, l, o) forecasting intervals. Here, colors indicate the kernel density estimation (KDE) of point concentration, with red representing high-density regions 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.

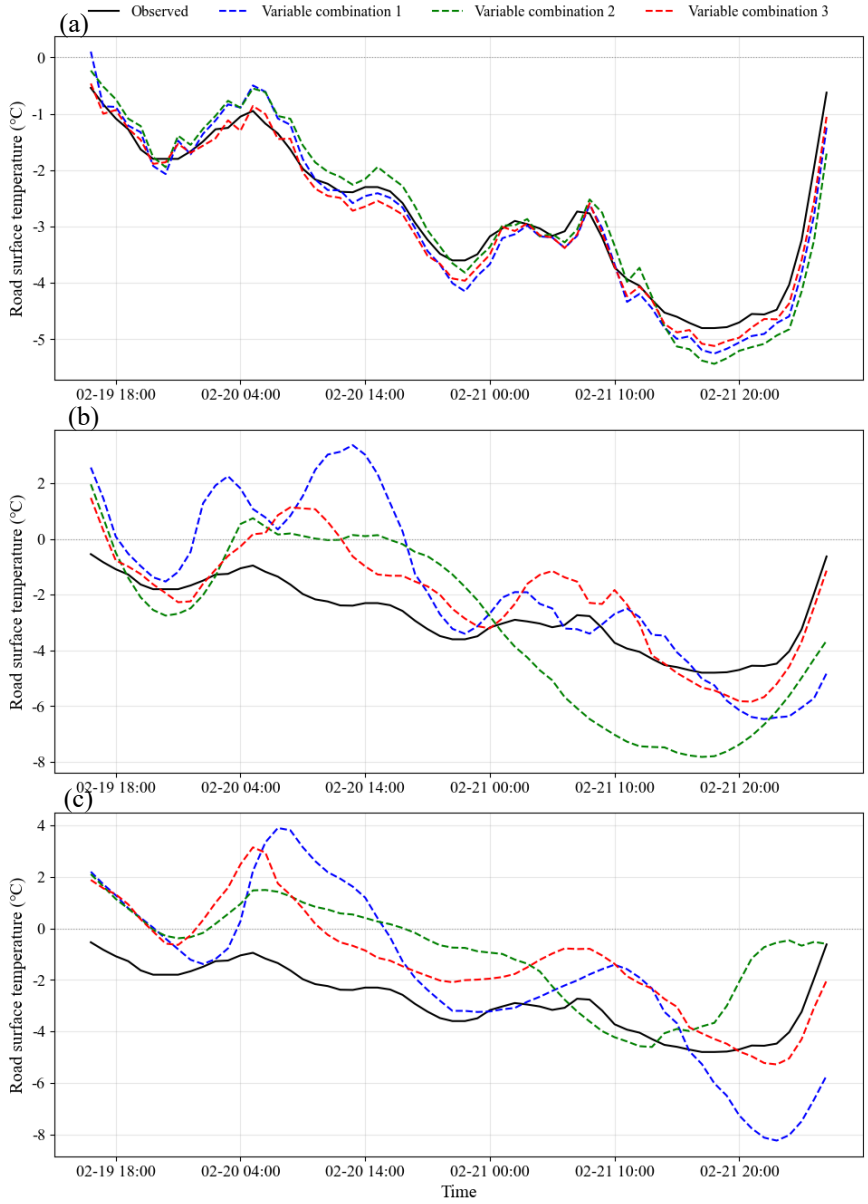


Figure 11. Winter RST prediction of ILES model with three input variable combinations in subzero low temperature period across 1-hour (a), 3-hour (b), and 6-hour (c) forecasting intervals.

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

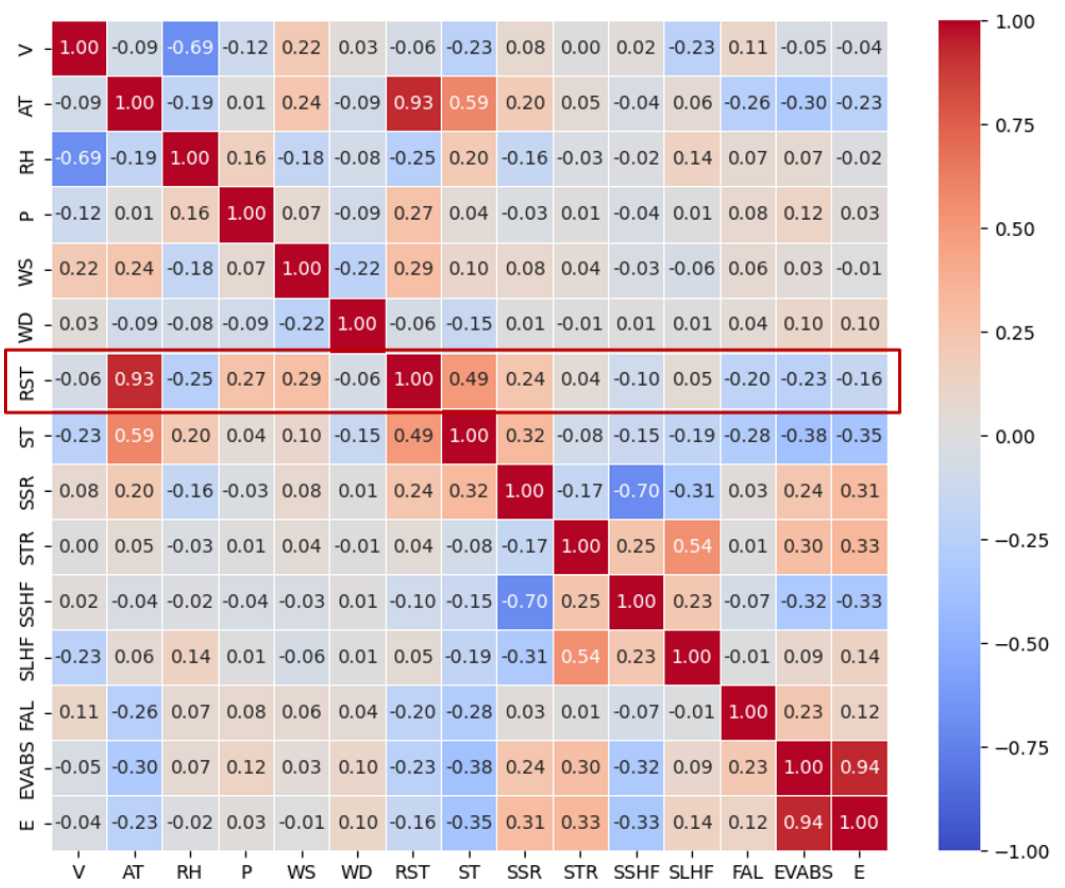


Figure 7. Spearman correlation coefficient plot between RST and various meteorological factors and ERA5_Land physical factors.

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.

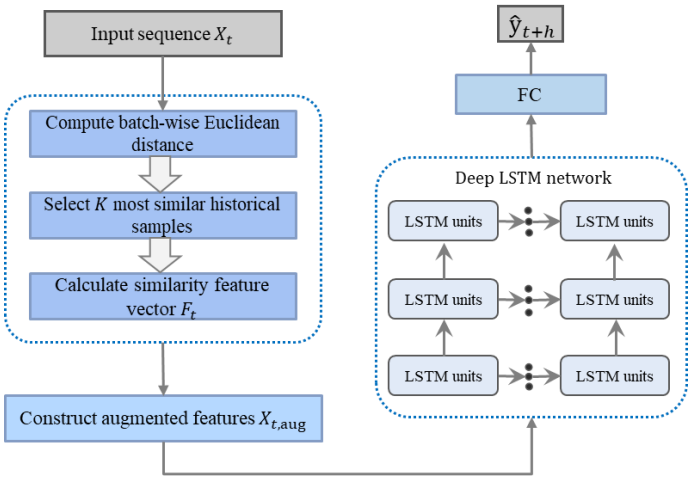


Figure S1. Architecture of the KNN-LSTM model.

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.

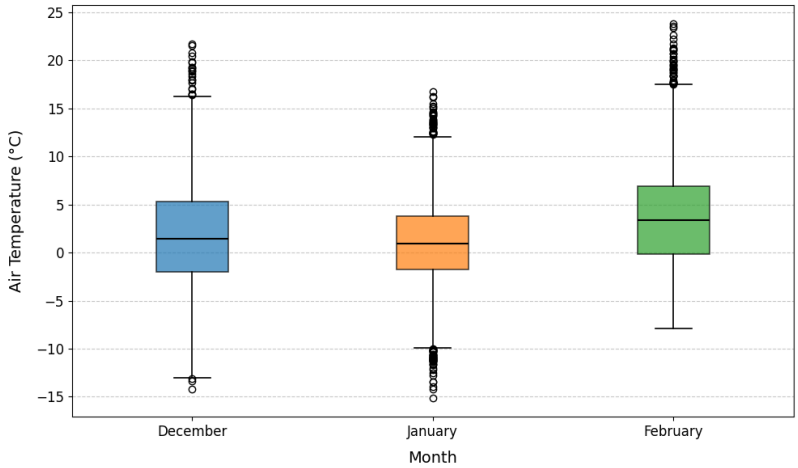


Figure S2. Monthly air temperature distribution.

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.

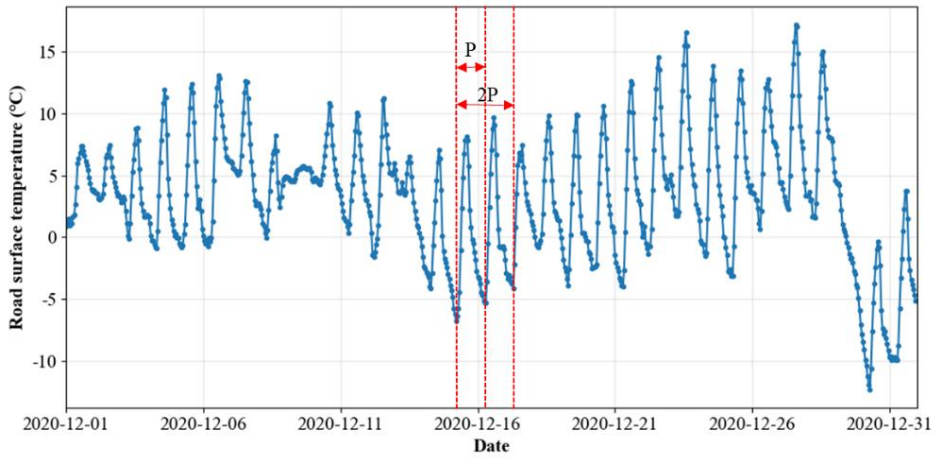


Figure 5. Periodic variation of RST. T represents a period, with time corresponding to one day.

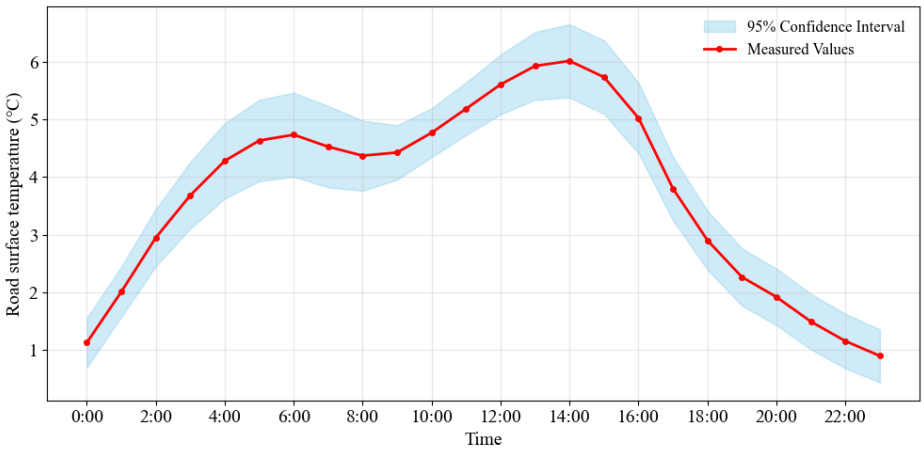


Figure 6. The 24-hour diurnal variation curve of RST. The shaded blue part of the graph is the 95% confidence interval for the RST.

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.