Dear Reviewer:

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

Major Comments

1. *"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;

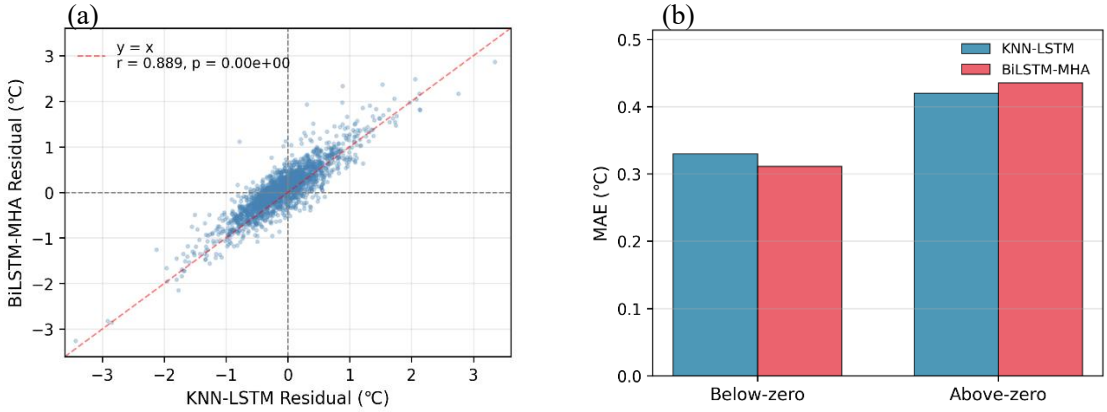
they constitute targeted technical innovations motivated by the specific challenges of RST time-series forecasting, and their empirical contributions are quantified in the ablation study (Table 4).

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.



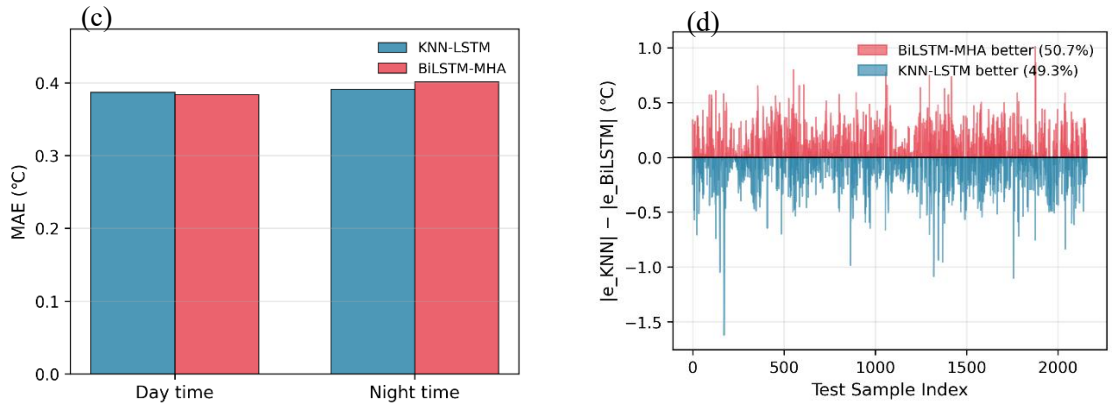


Figure 9. Complementarity analysis of base learners. Where (a) residual correlation analysis, (b) temperature interval error decomposition, (c) day and night time error decomposition, (d) time series advantage distribution.

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

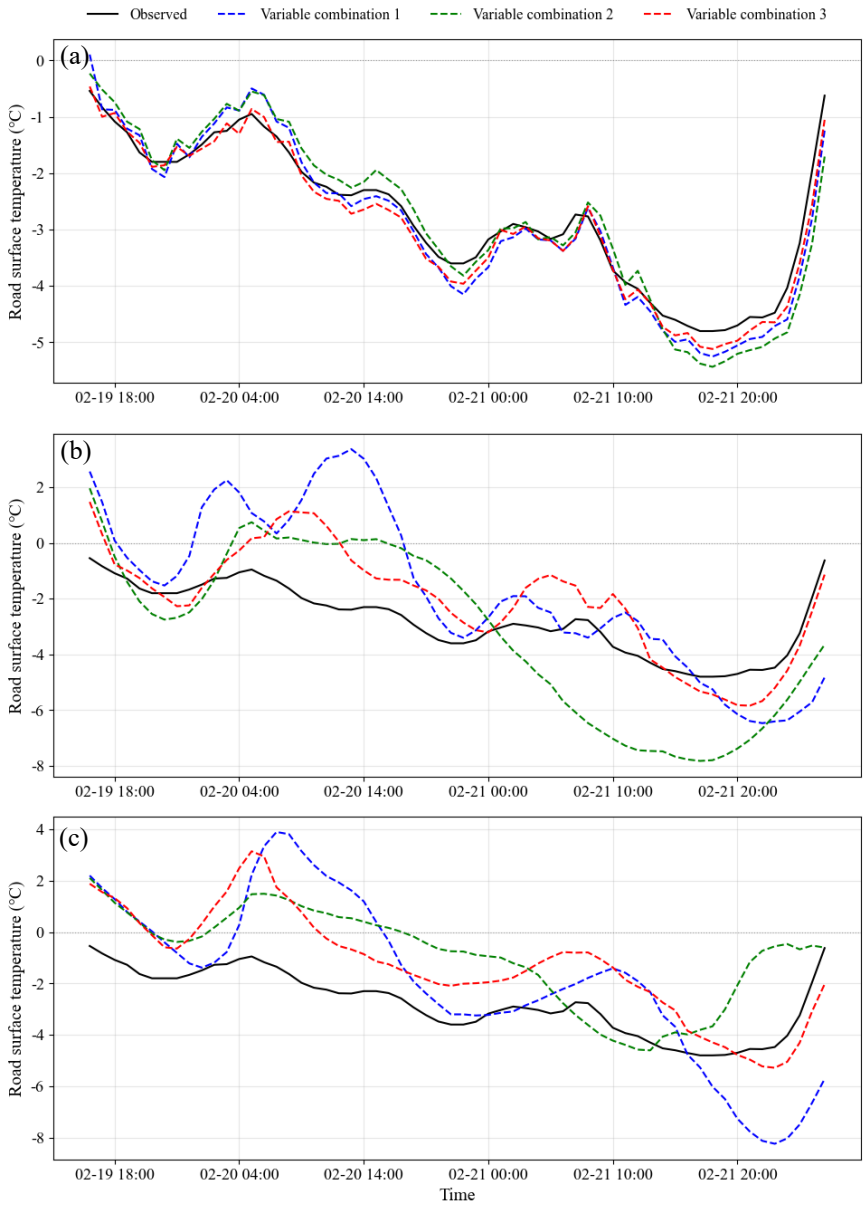
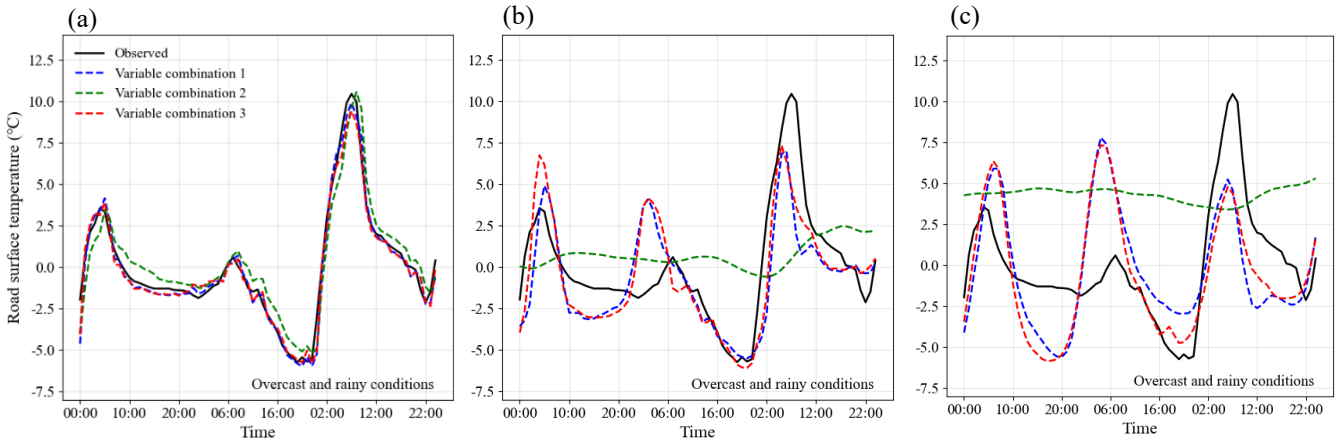


Figure 11. Winter RST prediction of ILES model with three input variable combinations in subzero low temperature period across 1-hour (a), 3-hour (b), and 6-hour (c) forecasting intervals. This period is the longest continuous section of the test sample with RST < 0 °C.



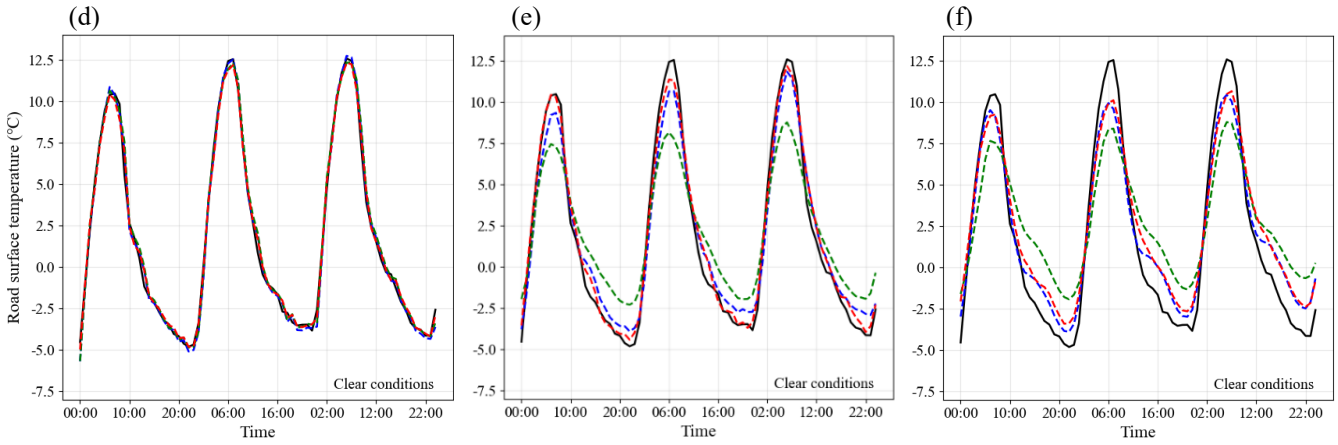


Figure 12. Winter RST prediction of ILES model with three input variables in overcast and rainy and clear synoptic conditions across 1-hour (a, d), 3-hour (b, e), and 6-hour (c, f) forecasting intervals.

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.

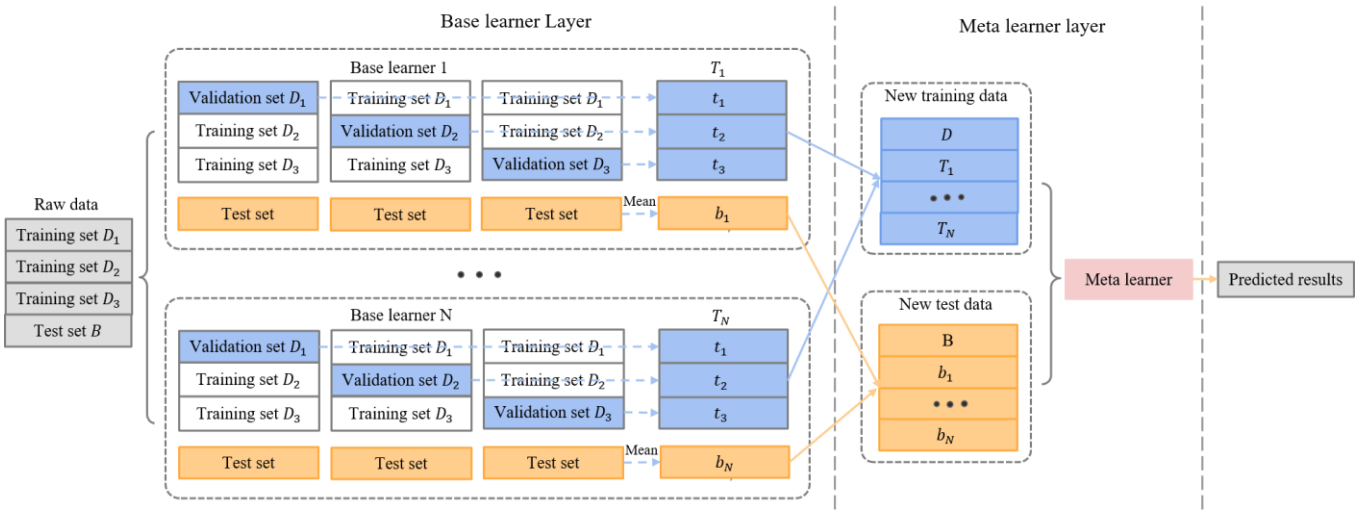


Figure 4. Architecture of the Stacking Ensemble based on Out-of-Fold Cross-Validation.

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.

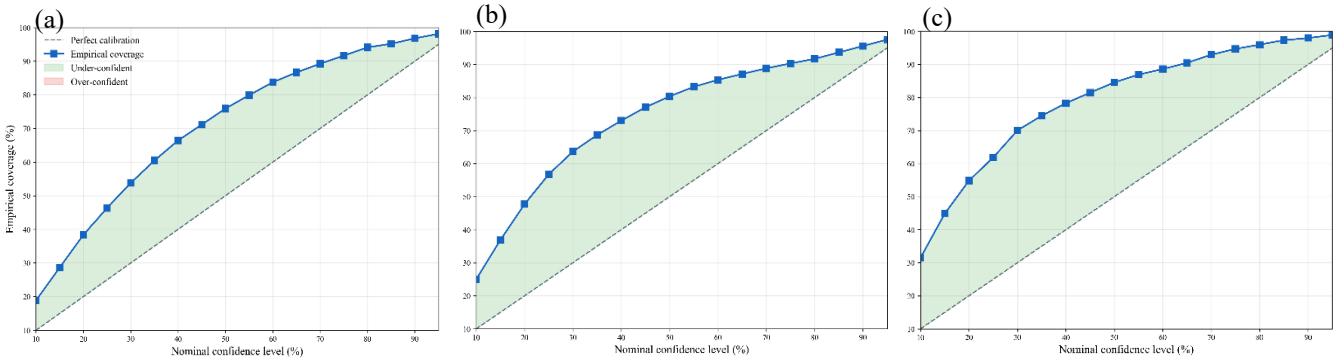


Figure 15. Reliability diagrams of the BRR meta-learner for 1-hour winter RST prediction at (a) M9393, (b) M9474, and (c) M9448.

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.

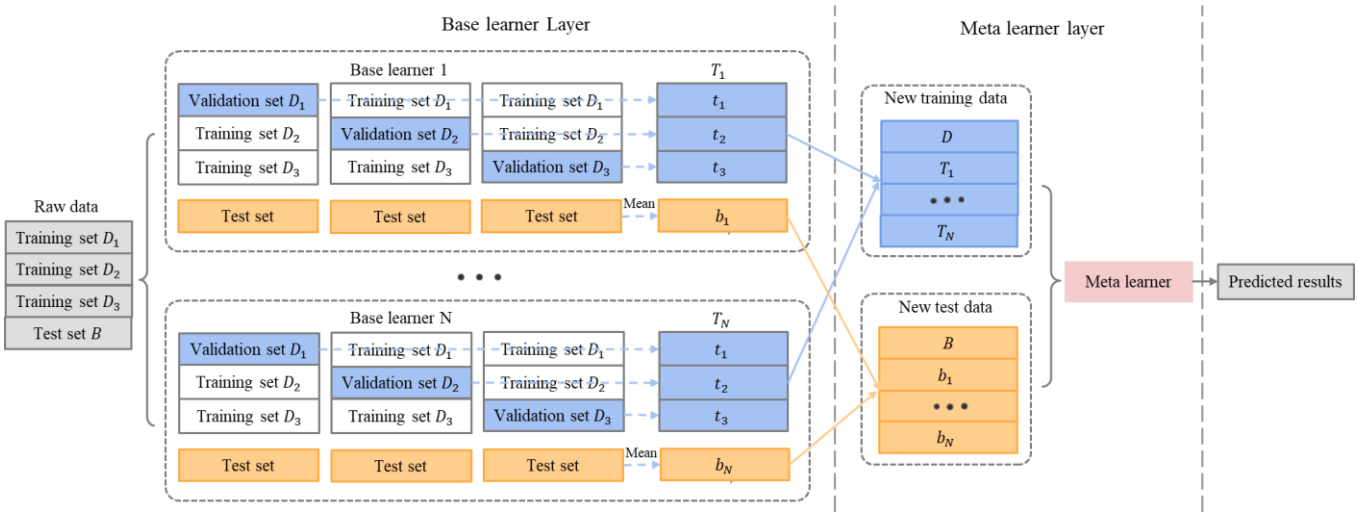


Figure 4. Architecture of the Stacking Ensemble based on Out-of-Fold Cross-Validation.

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.

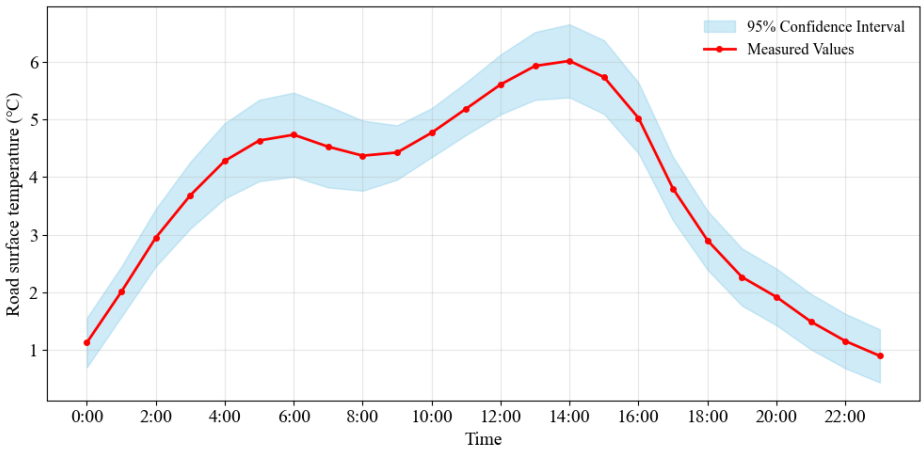


Figure 6. The 24-hour diurnal variation curve of RST. The shaded blue part of the graph is the 95% confidence interval for the RST.

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.

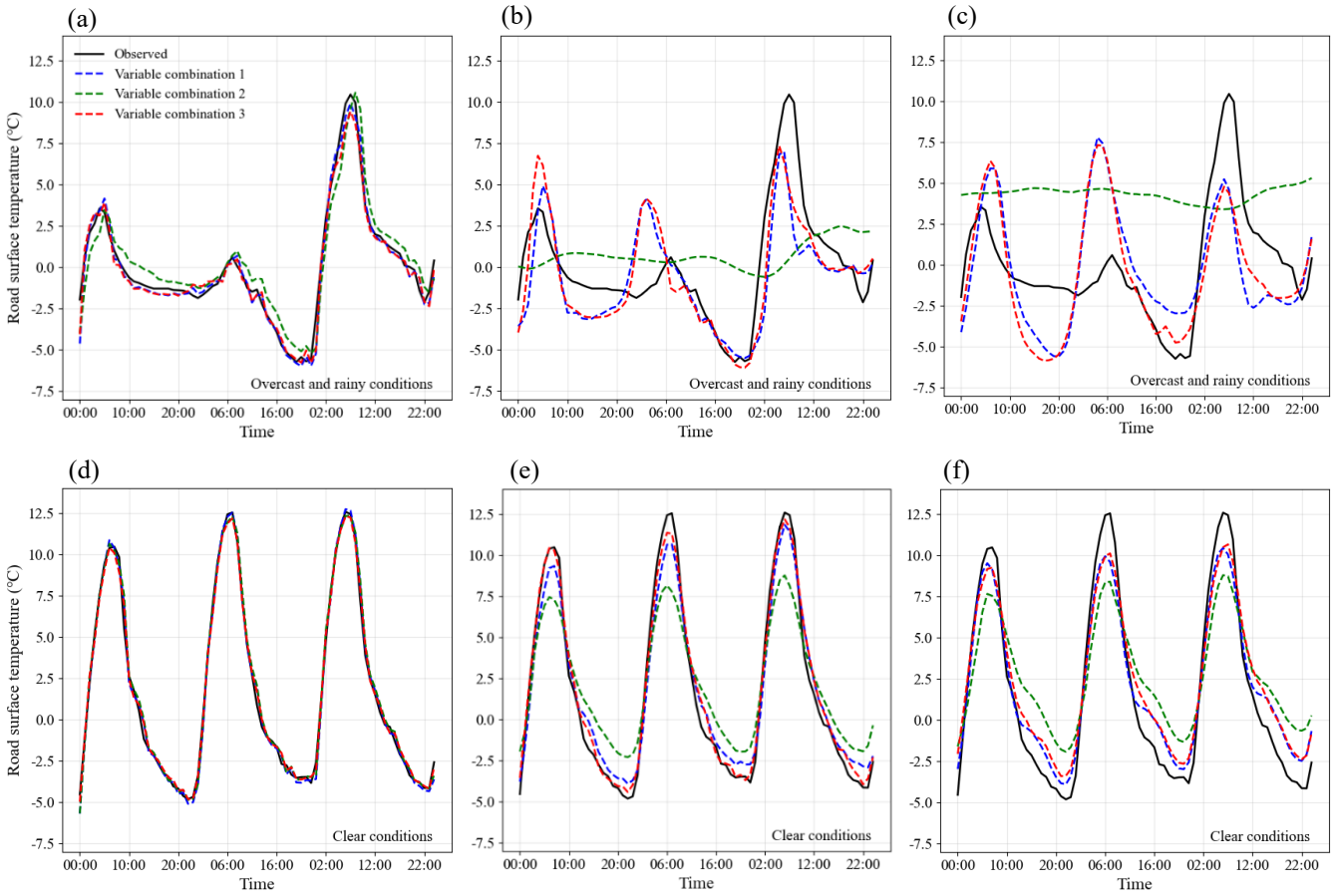


Figure 12. Winter RST prediction of ILES model with three input variables in overcast and rainy and clear synoptic conditions across 1-hour (a, d), 3-hour (b, e), and 6-hour (c, f) forecasting intervals.

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.

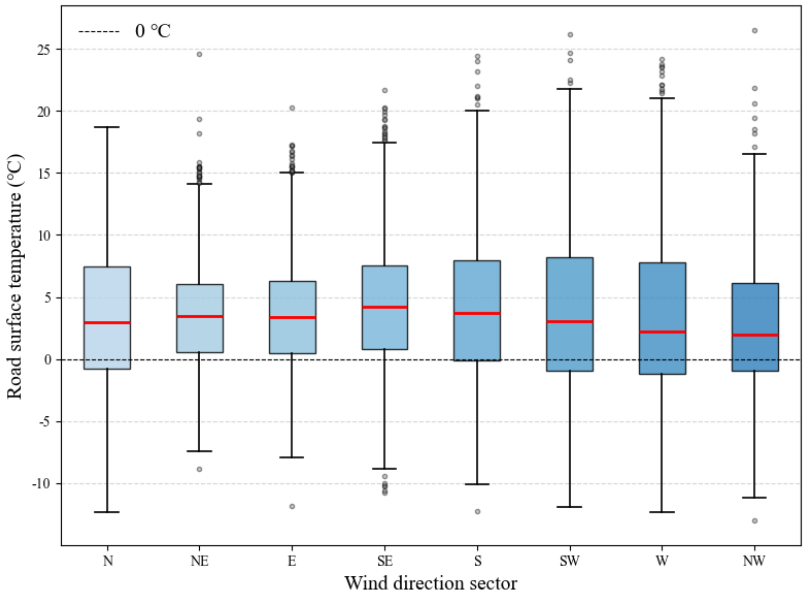


Figure S1. RST Distribution by Wind Direction Sector (8-sector classification, winter seasons 2020-2024).

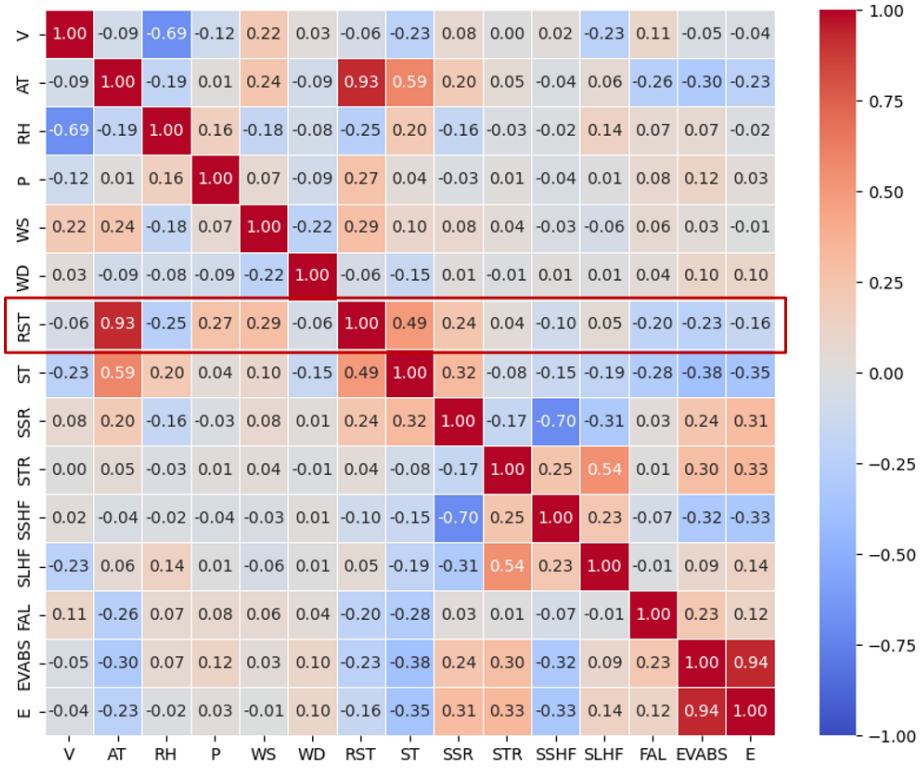


Figure 7. Spearman correlation coefficient plot between RST and various meteorological factors and ERA5_Land physical factors.

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.

References

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

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

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

Breiman, L.: Bagging predictors, Mach. Learn., 24(2), 123-140, https://doi.org/10.1007/BF00058655, 1996.

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

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

Darghiasi, P., Baral, A., Mattingly, S., et al.: Estimation of Road Surface Temperature Using NOAA Gridded Forecast Weather Data for Snowplow Operations Management, J. Cold Reg. Eng., 37(4), 04023018, https://doi.org/10.1061/JCREOE.0000686, 2023.

Darghiasi, P., Zamanian, M., and Shahandashti, M.: Enhancing Winter Maintenance Decision Making through Deep Learning-Based Road Surface Temperature Estimation, in Construction Res. Congr. 2024: Adv. Technol., Autom., and Comput. Appl. in Constr., ASCE, 701-711, https://doi.org/10.1061/9780784485262.70, 2024.

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

Diefenderfer, B. K., Al-Qadi, I. L., and Diefenderfer, S. D.: Model to predict pavement temperature profile: development and validation, J. Transp. Eng., 132(2), 162-167, https://doi.org/10.1061/(ASCE)0733-947X(2006)132:2(162), 2006.

Divina, F., Gilson, A., Goméz-Vela, F., et al.: Stacking ensemble learning for short-term electricity consumption forecasting, Energies, 11(4), 949, https://doi.org/10.3390/en11040949, 2018.

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

Gelman, A. and Shalizi, C. R.: Philosophy and the practice of Bayesian statistics, Br. J. Math. Stat. Psychol., 66(1), 8-38, https://doi.org/10.1111/j.2044-8317.2011.02037.x, 2013.

Hatamzad, M., Pinerez, G. C. P., and Casselgren, J.: Intelligent cost-effective winter road maintenance by predicting road surface temperature using machine learning techniques, Knowl.-Based Syst., 247, 108682, https://doi.org/10.1016/j.knosys.2022.108682, 2022.

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

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

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

Hoerl, A. E. and Kennard, R. W.: Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12(1), 55-67, https://doi.org/10.1080/00401706.1970.10488634, 1970.

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

Kingma, D. P.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, https://doi.org/10.48550/arXiv.1412.6980, 2014.

Kršmanc, R., Slak, A. Š., and Demšar, J.: Statistical approach for forecasting road surface temperature, Meteorol. Appl., 20(4), 439-446, https://doi.org/10.1002/met.1305, 2013.

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

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

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

MacKay, D. J. C.: Bayesian interpolation, Neural Comput., 4(3), 415-447, https://doi.org/10.1162/neco.1992.4.3.415, 1992.

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

Milad, A. A., Adwan, I., Majeed, S. A., et al.: Development of a hybrid machine learning model for asphalt pavement temperature prediction, IEEE Access, 9, 158041-158056, https://doi.org/10.1109/ACCESS.2021.3130010, 2021.

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

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

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

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

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

Schuster, M. and Paliwal, K. K.: Bidirectional recurrent neural networks, IEEE Trans. Signal Process., 45(11), 2673-2681, https://doi.org/10.1109/78.650093, 1997.

Spearman, C.: The proof and measurement of association between two things, in: Studies in Individual Differences: The Search for Intelligence, edited by: Jenkins, J. J. and Paterson, D. G., Appleton Century Crofts, 45-58, https://doi.org/10.1037/11491-005, 1961.

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

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

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

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

Yang, C. H., Yun, D. G., Kim, J. G., et al.: Machine learning approaches to estimate road surface temperature variation along road section in real-time for winter operation, Int. J. Intell. Transp. Syst. Res., 18(2), 343-355, https://doi.org/10.1007/s13177-019-00198-x, 2020.

Yin, Z., Hadzimustafic, J., Kann, A., et al.: On statistical nowcasting of road surface temperature, Meteorol. Appl., 26(1), 1-13, https://doi.org/10.1002/met.1737, 2019.

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