We sincerely thank the reviewer for the thorough evaluation and constructive suggestions. We have carefully addressed all comments and substantially revised the manuscript accordingly. Our point-by-point responses are detailed below, with corresponding changes highlighted in the revised manuscript.

Major Comments

(1) Insufficient Data Representativeness and Spatial Generalizability

Reviewer's Comment:

"The model is trained and validated exclusively on data from the Longhai Railway Bridge in Jiangsu, a region with a warm temperate semi-humid monsoon climate. This limits conclusions about performance in diverse geographies where RST dynamics differ due to terrain, vegetation, or pavement materials. The dataset lacks explicit coverage of extreme weather years (e.g., severe cold waves), raising questions about model reliability during rare but critical events."

Response:

We acknowledge this important limitation and have taken the following actions to address spatial generalizability:

Cross-site Validation Added: We have conducted additional validation using data from the Xiaohuangshan M9474 station, located on a cross-Yangtze River bridge in central Jiangsu Province (32.04°N, 119.86°E), representing different geographical and microclimatic conditions. The results are presented in Table 6, demonstrating that the ILES model consistently outperforms baseline models at both M9393 and M9474 sites, with 1-hour MAE of 0.373°C and 0.204°C respectively. This cross-site validation substantiates the model's transferability across geographically distinct locations.

Table 6: Comparison of MAE, RMSE, and MAPE for 1-hour winter pavement temperature prediction at M9393 and M9474 sites in 2024 using five deep learning models (two foundation models LSTM and BiLSTM, two improved hybrid models KNN-LSTM and BiLSTM-MHA, and the integrated model ILES proposed in this paper).

Model	M9393			M9474		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
LSTM	0.453	0.636	46.90	0.229	0.330	5.26
BiLSTM	0.422	0.572	42.83	0.228	0.358	4.85
KNN-LSTM	0.389	0.536	38.08	0.218	0.331	5.46
BiLSTM-MHA	0.393	0.545	41.40	0.228	0.347	5.89
ILES	0.373	0.521	37.97	0.204	0.322	5.27

Extreme Weather Coverage: While our dataset spans four winter periods (2020-2024), capturing substantial temperature variability (ranging from -15.15°C to 23.79°C for air temperature and -12.99°C to 26.53°C for RST as shown in Table 2), we acknowledge the dataset includes 1,338 sub-zero temperature samples (RST < 0°C), representing approximately 15.4% of the total test set. Section 4.2 specifically evaluates model performance under these critical low-temperature conditions, demonstrating that ILES achieves MAE of 0.292 °C for 1-hour forecasts under sub-zero regimes, representing 23.2% improvement over baseline LSTM. However, we acknowledge in the revised Conclusion (Section 5) that future research should incorporate data from additional stations across diverse climatic zones to further enhance spatial representativeness and validate performance under more extreme meteorological conditions.

Manuscript Revisions:

1. Added Section 4.2 with dedicated analysis of sub-zero temperature prediction.
2. Added Table 6 presenting cross-site validation results at M9474 station.
3. Expanded Conclusion to acknowledge limitations and propose multi-site validation as future work.
4. Clarified data coverage in Section 3.1, explicitly stating the temperature ranges and extreme weather representation.

(2) Integrate Meteorological Physics to Enhance Interpretability

Reviewer's Comment:

"Incorporate key parameters from the road surface energy balance equation (e.g., albedo, thermal conductivity, estimated solar radiation) as inputs or constraints."

Response:

We appreciate this valuable suggestion to strengthen the physical basis of our model. We acknowledge that the variables

mentioned by the reviewer—albedo, thermal conductivity, and solar radiation—indeed influence road surface temperature through their roles in the surface energy balance equation. According to existing literature on pavement temperature prediction, meteorological factors affecting road surface temperature can be broadly categorized into several primary groups: thermal drivers, atmospheric state parameters and precipitation-related variables. Thermal drivers encompass solar radiation and air temperature, the former serving as the fundamental energy source for pavement heating and the latter governing the convective heat exchange between the road surface and the surrounding air. Atmospheric state parameters include relative humidity and wind speed, both of which modulate the rate of evaporative cooling and conductive heat transfer at the pavement–air interface. Precipitation-related variables refer to precipitation intensity and duration, factors that directly alter the pavement's thermal properties through processes such as water absorption and latent heat exchange (Chen et al., 2019; Gui et al., 2007; Krsmanć et al., 2013; Liu et al., 2018; Zhang et al., 2024; Stoner et al., 2019)

Regarding solar radiation, we explicitly address this in Section 3.2: "Although solar radiation determines the heat distribution of road surfaces, it mainly affects road temperatures during the daytime and is absent at night (Qin et al., 2022). In addition, constraints associated with data availability precluded the incorporation of this variable in the present study." Our model captures diurnal thermal cycles through lagged RST inputs (Figures 5-6), which implicitly encode solar radiation effects through the observed temperature periodicity. The correlation analysis (Figure 7) guided selection of five physically relevant predictors: air temperature (heat conduction), relative humidity (evaporative cooling), precipitation (thermal properties), wind speed (convective heat transfer), and historical RST (thermal inertia).

We have incorporated SHAP (SHapley Additive exPlanations) analysis in Section 4.3 to quantify the physical contribution of each meteorological variable to RST predictions (Joo et al., 2023). As presented in Figure 14, air temperature emerges as the dominant feature contributing 38.09% of model importance, consistent with heat conduction principles. Wind speed, relative humidity, and precipitation contribute 22.23%, 20.70%, and 18.99% respectively. Notably, under precipitation events as illustrated in Fig. 14(b,d), the contributions of humidity and precipitation rise markedly, with the former reaching 22.67% and the latter accounting for 19.21%. This variation provides a mechanistic interpretation for the heightened prediction uncertainty documented during such periods, thereby validating the capacity of the proposed model to dynamically assign weights to distinct information sources in accordance with prevailing meteorological condition

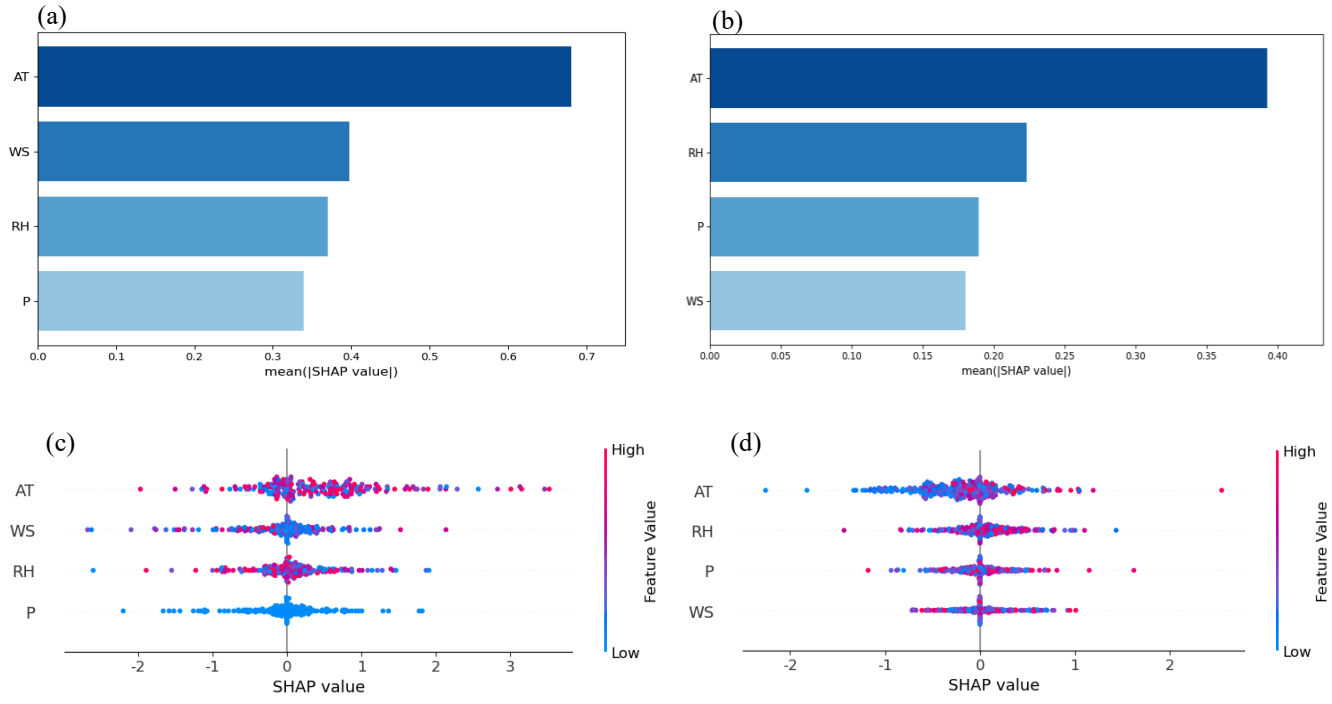

Figure 14: Global (a, c) and precipitation event (b, d) feature importance of variables and SHAP results.

While direct integration of albedo and thermal conductivity as model inputs is challenging due to their site-specific and time-invariant nature in our single-site dataset, the BiLSTM-MHA architecture with multi-head attention mechanisms (Section 2.2) enables the model to learn implicit representations of energy balance dynamics. The attention mechanism dynamically

weights temporal features corresponding to different physical processes (short-term fluctuations from convective cooling, long-term trends from thermal inertia, periodic patterns from solar forcing), as discussed in relation to Equation 10.

$$\mathrm{MuItiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)\mathrm{W}^O\ , \quad (10)$$

Manuscript Revisions:

1. Added Section 4.3 with comprehensive SHAP analysis (Figures 14-15).
2. Expanded Section 3.2 to justify feature selection based on physical mechanisms.
3. Enhanced discussion of how model architecture captures energy balance dynamics.
4. Acknowledged in Conclusion that future work should explore explicit integration of physical constraints through hybrid physics-informed neural networks.

(3) Inadequate Benchmarking Against State-of-the-Art Models

Reviewer's Comment:

"The manuscript claims superiority over 'individual models' (LSTM, KNN-LSTM, Attention-BiLSTM) but lacks comparisons with recent hybrid methods in RST prediction. Quantitative metrics (e.g., MAE, MSE) against these models are absent, weakening claims of methodological advancement."

Response:

We have substantially expanded the benchmarking analysis to address this concern:

Comprehensive Model Comparison: Table 4 now presents systematic evaluation of eight models: (1) traditional machine learning methods RF and XGBoost (Darghiasi et al., 2025; Kebede et al., 2024), (2) deep learning baselines LSTM and BiLSTM, (3) recent hybrid architecture CNN-LSTM (Tabrizi et al., 2021), (4) our proposed base learners KNN-LSTM and BiLSTM-MHA, and (5) the ILES ensemble model. This comparison spans three forecasting horizons (1-hour, 3-hour, 6-hour) across four metrics (MAE, RMSE, MAPE, R^2).

Table 4: Performance evaluation across 1-, 3-, and 6-hour forecasting intervals.

Model	1-hour			3-hour			6-hour		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
RF	0.524	0.780	45.11	1.416	1.976	128.86	2.261	3.293	212.46
XGB	0.571	0.893	41.94	1.357	1.887	129.54	2.159	3.146	211.85
LSTM	0.453	0.636	46.90	1.453	2.198	130.01	2.366	3.452	251.24
BiLSTM	0.422	0.572	42.83	1.414	2.056	144.70	2.252	3.092	254.60
CNN-LSTM	0.406	0.567	39.46	1.399	1.399	157.10	2.154	2.921	235.35
KNN-LSTM	0.389	0.536	38.08	1.285	1.927	114.25	2.259	2.923	227.92
BiLSTM-MHA	0.393	0.545	41.40	1.517	1.993	129.88	2.194	2.822	275.45
ILES	0.373	0.521	37.97	1.291	1.808	113.38	2.094	2.688	243.97

Quantitative Performance Gains: Section 4.1 provides detailed quantitative comparisons:

For 1-hour forecasting, ILES achieves MAE of 0.373°C, demonstrating 4.11% improvement over KNN-LSTM (0.389°C), 5.09% over BiLSTM-MHA (0.393°C), 8.13% over CNN-LSTM (0.406°C), 11.6% over BiLSTM (0.422°C), and 17.7% over baseline LSTM (0.453°C).

ILES achieves 28.82% lower MAE compared to RF (0.524°C) and 34.68% lower MAE compared to XGBoost (0.571°C).

At 6-hour horizon, ILES (MAE: 2.094°C) maintains 7.30% advantage over KNN-LSTM and 4.56% over BiLSTM-MHA, with substantially greater improvements over baseline methods (11.5% better than LSTM, 7.02% better than BiLSTM).

Literature Contextualization: Table 1 now provides systematic comparison with recent studies, positioning our work against state-of-the-art methods including CNN-LSTM (Tabrizi et al., 2021), Attention-BiLSTM (Bai et al., 2022), GRU/LSTM

(Dai et al., 2023), and RF-LSTM (Zhang et al., 2024), with explicit comparison of metrics, features, and dataset characteristics.

Table 1: Comparison of different RST prediction models.

References	Model	Metrics	Features	Data sizes	Time interval
Tabrizi et al. (2021)	CNN-LSTM	MAE, RMSE, MAPE, R^2, NSE	RST, AT, Year, Month	10895	1h
Milad et al. (2021)	Bi-LSTM	MAE, MSE, MAPE, R^2	RST, AT, Depth, Time	7200	1h
Bai et al. (2022)	Att-BiLSTM	MAE, MSE, MAPE	RST, V, AT, RH, WD, WS, P	4344	1h
Dai et al. (2023)	GRU, LSTM	MAE, MSE, MAPE	RST, V, AT, RH, P, WS, WD	8640	1h
Zhang et al. (2024)	RF-LSTM	MAE, MSE, RMSE, MAPE	RST, AT, RH, WS, WD, P et al.	-	10min
Our paper	ILES	MAE, RMSE, MAPE, R^2	RST, V, AT, RH, P, WS, WD	8664	1h

Manuscript Revisions:

1. Expanded Table 4 to include RF, XGBoost, and CNN-LSTM benchmarks.
2. Added comprehensive quantitative analysis in Section 4.1 with percentage improvements.
3. Added Figures 9-10 for statistical comparison of error distributions.
4. Enhanced Table 1 with detailed literature comparison.
5. Strengthened discussion of performance gains relative to each baseline method.

Minor Comments

(1) Ambiguities in Figure and Table Presentations

Reviewer's Comment:

"Figure 6 (RST periodicity) lacks confidence intervals, making it impossible to assess the statistical significance of diurnal variations."

Response:

We have revised Figure 6 to include 95% confidence intervals for the 24-hour diurnal variation curve. The shaded blue region now clearly illustrates the degree of dispersion in temperature variations across different dates at the same hour, demonstrating statistically significant periodicity. The accompanying text in Section 3.2 has been updated to describe: "The 95% confidence interval reflects the degree of dispersion in temperature variations across different dates at the same hour."

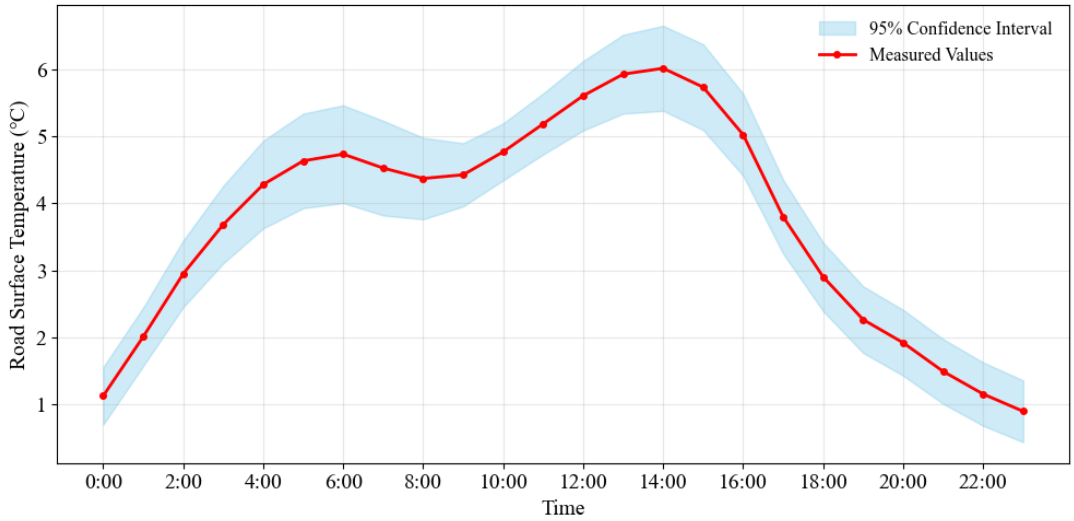

Figure 6: The 24-hour diurnal variation curve of RST. The shaded blue part of the graph is the 95% confidence interval for the RST.

Manuscript Revisions:

1. Updated Figure 6 with 95% confidence intervals.

2. Enhanced caption and Section 3.2 text to interpret confidence intervals.

(2) Inconsistent Terminology and Citation Errors

Reviewer's Comment:

"The term 'Attention-LSTM' is used in Figure 12's caption but not defined in the main text; it should be corrected to 'Attention-BiLSTM' for consistency. In Section 4.2, 'STM' is referenced in Figure 12's legend but not defined, causing confusion."

Response:

We sincerely apologize for these inconsistencies. We have systematically corrected all terminology throughout the manuscript:

Terminology Standardization:

a) Unified model naming: "BiLSTM-MHA" (Bidirectional LSTM with Multi-Head Attention) is used consistently throughout the manuscript (Abstract, Introduction, Methodology, Results, Figures, Tables).

b) Ensemble model consistently referred to as "ILES" (Improved LSTM Ensemble with Stacking).

c) All figure captions and legends updated to match standardized terminology.

Definition Clarifications:

a) Section 2.2 now clearly defines the BiLSTM-MHA architecture with explicit explanation of the multi-head attention mechanism (Equations 10-14).

b) Figure 3 caption provides comprehensive notation definitions.

c) All undefined abbreviations have been removed or properly defined at first use.

Manuscript Revisions:

1. Corrected all instances of inconsistent model naming throughout manuscript.
2. Verified consistency across Abstract, main text, figures, tables, and captions.
3. Added explicit model name definitions in Section 2.

Summary of Key Improvements

In response to the reviewer's concerns, we have made the following major enhancements:

1. Enhanced Generalizability: Added cross-site validation (Table 6) and expanded discussion of extreme weather coverage.
2. Strengthened Physical Interpretability: Incorporated SHAP analysis (Section 4.3, Figures 14-15) to quantify physical variable contributions.
3. Comprehensive Benchmarking: Expanded comparison to eight models including RF, XGBoost, CNN-LSTM with detailed quantitative analysis (Table 4, Figures 9-10).
4. Improved Presentation Quality: Added confidence intervals to Figure 6, standardized terminology, and corrected all inconsistencies.
5. Expanded Discussion: Enhanced Introduction and Conclusion to acknowledge limitations and propose future research directions.

We believe these revisions substantially strengthen the manuscript's scientific rigor, reproducibility, and practical relevance for operational road weather management. We are grateful for the reviewer's constructive feedback, which has significantly improved the quality of our work.