A Hybrid Method for Winter Road Surface Temperature Prediction Using Improved LSTMs and Stacking-Based Ensemble Learning

Wanting Li[1,2,3], Linyi Zhou[4,5], Xianghua Wu[1,2,3], Yuanhong Guan[1,2,3], Yuanhao Guo[1], Kun Chen[1], Huiwen Lin[1]

[1] School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] Center for Applied Mathematics of Jiangsu Province, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] Jiangsu International Joint Laboratory on System Modeling and Data Analysis, Nanjing University of Information Science and Technology, Nanjing 210044, China
[4] Nanjing Innovation Institute for Atmospheric Sciences, Chinese Academy of Meteorological Sciences–Jiangsu Meteorological Service, Nanjing 210041, China
[5] Jiangsu Key Laboratory of Transportation Meteorology of CMA / Key Laboratory of Severe Storm Disaster Risk, Nanjing 210041, China

Correspondence to: Linyi Zhou (zhoulinyi@cma.gov.cn), Xianghua Wu (wuxianghua@nuist.edu.cn)

Abstract. Accurate prediction of road surface temperature (RST) is essential for proactive winter road maintenance and traffic safety management. However, existing approaches—ranging from physics-based models to data-driven methods—either require detailed pavement thermal parameters that are rarely available at operational road meteorological stations, or lack the capacity to simultaneously exploit local meteorological analogues and long-range temporal dependencies in an interpretable ensemble framework. This study proposes the Improved LSTMs Ensemble with Stacking (ILES) framework, which integrates two base learners with partially complementary predictive characteristics~~two architecturally complementary base learners~~ within a ~~leakage-free~~ stacking ensemble employing out-of-fold cross-validation to prevent meta-learner training leakage. The first base learner, KNN-LSTM, augments sequential ~~modeling~~modelling with similarity-based retrieval of historically analogous meteorological states to capture locally recurrent patterns. The second, BiLSTM-MHA, combines bidirectional recurrent processing with multi-head self-attention to extract long-range temporal dependencies across a 24-hour input window. Moreover, a Bayesian Ridge Regression meta-learner fuses the base-learner outputs through Evidence Maximization, yielding probabilistic forecasts with closed-form posterior predictive uncertainty at the ensemble combination layer~~a Bayesian Ridge Regression meta-learner fuses the base learner outputs through Evidence Maximization, yielding probabilistic forecasts with closed-form uncertainty quantification~~. The framework is trained and evaluated on four consecutive winter seasons (December 2020 to February 2024) of road meteorological observations from station M9393 in the northwest inland plain of Jiangsu Province, China. Results indicate that ILES achieves lower prediction errors in general than ten other models spanning persistence forecasting, traditional machine learning, and deep learning approaches, with R^2 reaching 0.993, 0.923, and 0.826 at 1-, 3-, and 6-hour forecasting horizons, respectively. Among three input configurations evaluated, physics-motivated feature engineering incorporating the air–surface temperature gradient and multi-scale RST temporal tendencies

~~outperforms~~outperform both the station-only baseline and ERA5-Land reanalysis augmentation, indicating that domain-knowledge-guided feature construction provides a more effective and operationally practical input strategy at instrumented sites. SHAP-based interpretability analysis, stratified by temperature regime and diurnal cycle, confirms that the learned feature importance rankings are qualitatively consistent with the dominant drivers of RST evolution identified by surface energy balance theory. Multi-site generalization is further validated at two independent stations within the same temperate monsoon climate zone, confirming the transferability of the proposed framework across different road environments within this climate setting~~Multi-site generalization is further validated at two independent stations, confirming the transferability of the proposed framework across different road environments~~.

1 Introduction

Accurate prediction of road surface temperature (RST) is a prerequisite for effective winter road maintenance and traffic-safety management (Shao and Lister, 1996; Nantasai et al., 2019; Zhao et al., 2020). Sub-zero pavement temperatures substantially reduce the pavement friction coefficient, increasing the risk of vehicle accidents and network-wide congestion. Because RST governs the phase state of moisture at the pavement-atmosphere interface, it serves as the primary indicator for determining whether wet road surfaces will freeze under prevailing meteorological conditions (Li et al., 2022; Nowrin and Kwon, 2022; Zhang et al., 2023).

From a physical standpoint, RST is governed by the surface energy balance (SEB) at the pavement–atmosphere interface. Following Hermansson (2004) and Chen et al. (2019), the net energy flux at the road surface can be expressed as:

$$Q^* = Q_h + Q_e + Q_G + Q_m \,, \tag{1}$$

where Q^* is the net radiation flux; Q_h is the turbulent sensible heat flux driven by the surface–air temperature gradient and wind speed; Q_e is the latent heat flux modulated by relative humidity and precipitation; Q_G is the pavement heat storage flux governed by the substrate's thermal conductivity and heat capacity; and Q_m represents the latent heat of fusion associated with freezing or melting of surface moisture. All flux terms adopt a sign convention in which positive values denote energy directed into the road surface and negative values denote energy loss from the surface; under this convention, Q^* represents the net radiative energy available to drive the remaining flux terms. RST reflects the cumulative outcome of these fluxes over preceding hours, exhibiting diurnal periodicity, lagged responses to radiative forcing, and abrupt thermal transitions during precipitation events or cold-air advection (Hermansson, 2004; Chen et al., 2019). This physical understanding motivates both the selection of meteorological predictors and the design of derived input features in the present study.

RST prediction presents two fundamental challenges that existing approaches have not simultaneously resolved. Physics-based models—including finite-difference, finite-element, and finite-volume formulations—solve the surface energy balance equation coupled to one-dimensional heat conduction in the pavement substrate and can yield physically interpretable, high-accuracy solutions when site-specific thermal parameters are well characterised (Hermansson, 2004; Wang et al., 2009; Chen et al., 2019; Minhoto et al., 2005; Schindler et al., 2004; Saliko et al., 2023). However, their accuracy is critically sensitive to pavement thermal and radiative properties—thermal conductivity, volumetric heat capacity, surface emissivity, and the convective heat transfer coefficient—that are rarely available at operational road meteorological stations (Qin and Hiller, 2013; Athukorallage et al., 2023; Ayasrah et al., 2023; Adwan et al., 2021). This parameter identifiability constraint extends to operational systems such as METRo (Crevier and Delage, 2001), whose performance remains dependent on accurate specification of pavement layer properties, surface albedo, and anthropogenic heat sources that vary across road segments and degrade over time (Shao and Lister, 1996; Kangas et al., 2015; Karsisto et al., 2016), fundamentally limiting the scalability of physics-based approaches to the large number of heterogeneous road segments for which detailed subsurface characterisation is unavailable. Concurrently, RST is a strongly autocorrelated variable whose evolution reflects the integrated effect of meteorological forcing over preceding hours, exhibiting lagged responses to radiative and turbulent fluxes and regime-dependent nonlinear dynamics during precipitation events or nocturnal radiative cooling. Statistical and empirical models—including multiple linear regression, stepwise regression, and regression-based nowcast systems—have identified near-surface air temperature and lagged RST as the dominant predictors (Yin et al., 2019; Kršmanc et al., 2013; Li et al., 2018; Diefenderfer et al., 2006; Asefzadeh et al., 2017; Hassan et al., 2005), yet these approaches impose linearity or low-order polynomial constraints on the inherently nonlinear meteorology–RST relationship, precluding adequate representation of the multi-hour temporal dependencies and abrupt thermal transitions that characterise real-world RST evolution (Wang, 2015; Gedafa et al., 2014; Darghiasi et al., 2025; Jing and Zhang, 2018).

Data-driven machine learning (ML) methods address the nonlinearity limitation by learning flexible input–output mappings directly from observational records without requiring explicit thermal parameter calibration. Ensemble tree methods, including random forest (RF), gradient boosting decision trees (GBDT), and extreme gradient boosting (XGBoost), have achieved strong predictive performance by aggregating predictions from multiple weak learners (Milad et al., 2021b; Qiu et al., 2020; Liu et al., 2018; Kebede et al., 2024; Yuan et al., 2025). Support vector regression (SVR) and Gaussian process regression (GPR) have captured nonlinear RST–meteorology relationships under varying surface conditions (Molavi et al., 2022; Wang et al., 2023), and artificial neural networks (ANNs) have demonstrated superior pattern-learning capacity relative to conventional regression approaches (Abo-Hashema, 2013; Rigabadi et al., 2022). However, these methods treat each input sample independently and therefore cannot exploit the multi-hour temporal autocorrelation structure inherent in RST time series, constraining their ability to represent the thermal inertia and lag dynamics that are characteristic of RST evolution (Yang et al., 2020; Hatamzad et al., 2022; Darghiasi et al., 2024).

Recurrent deep learning architectures, and in particular Long Short-Term Memory (LSTM) networks and their variants, address this limitation more directly by learning sequential dependencies through gating mechanisms that selectively retain and discard information over multi-hour input windows (Dai et al., 2023; Ghalandari et al., 2023). Several studies have demonstrated their superiority over traditional ML approaches for RST prediction: Tabrizi et al. (2021) showed that a hybrid CNN-LSTM model outperformed LSTM, ConvLSTM, Seq2Seq, and WaveNet baselines across 1-, 2-, 4-, and 6-hour horizons; Zhang et al. (2023) reported MAE of 0.82°C and RMSE of 1.24°C with an XGBoost-LSTNet combination; Li et al. (2022) showed that a hybrid Bayesian structural time series-Bayesian neural network model attained greater than 95% coverage within predicted confidence intervals; Dai et al. (2023) proposed a GRU-LSTM ensemble exploiting RST periodicity and meteorological lag effects; and Zhang et al. (2024) integrated RF-based feature selection with LSTM for short-term prediction at 10-minute resolution. Bidirectional LSTM (BiLSTM) architectures, which process the input sequence in both forward and backward temporal directions, have shown enhanced feature-extraction capabilities relative to unidirectional LSTMs (Maddu et al., 2021; Tao et al., 2024; Milad et al., 2021a). Bai et al. (2022) demonstrated that an attention-based Bi-LSTM achieved 93.4% of predictions within 1°C error. Table 1 ~~summarises~~summarizes recent LSTM-based approaches for RST prediction, including the model architectures employed, the input features considered, and representative predictive accuracy reported in each study.

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

Reference	Model	Feature	Interval	Characteristic	Evaluation
Tabrizi et al. (2021)	CNN-LSTM	RST, AT, SR	1h	CNN-LSTM hybrid architecture for multi-horizon forecasting	MAE=1.05–3.43 °C and R^2=0.98–0.80 across 1- to 6-hour horizons
Milad et al. (2021a)	Bi-LSTM	AT, Depth, Time	1h	Bidirectional LSTM with enhanced feature extraction across depth and time dimensions	MAE = 1.332 °C; R^2 = 0.956
Bai et al. (2022)	Att-BiLSTM	RST, AT, P, WS, RH	1h	Attention-augmented BiLSTM with sliding window ~~optimisation~~optimization 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;~~ $\cancel{R^2 = 0.993\text{–}0.826}$

Despite these advances, the existing literature exhibits several limitations that warrant further investigation. First, individual LSTM-based architectures tend to emphasise either local pattern recurrence associated with historically similar meteorological trajectories or long-range temporal dependencies spanning the full input window, but rarely capture both simultaneously. While some studies have explored ensemble combinations of heterogeneous models, the systematic integration of architecturally complementary LSTM variants through principled meta-learning remains underexplored for RST prediction. Moreover, the consistency of ensemble-derived predictive gains across multiple forecasting horizons has not been thoroughly demonstrated. Second, the relative contribution of different input data sources to RST prediction accuracy remains insufficiently characterised. While several studies have incorporated reanalysis products or derived meteorological variables alongside station observations, systematic comparisons across input configurations under consistent experimental conditions are scarce, making it difficult to determine whether the additional data complexity is justified by commensurate predictive gains. Third, although post-hoc feature attribution methods such as SHAP (SHapley Additive exPlanations; Lundberg and Lee, 2017) have been applied to data-driven RST models, existing analyses have typically reported globally averaged attribution values without systematic stratification by temperature regime or time of day. Such stratified analysis would provide a more rigorous basis for evaluating whether learned model ~~behavior~~behaviour is qualitatively consistent with the dominant meteorological drivers identified by SEB theory, and for identifying conditions under which the model may be less reliable.

Motivated by these limitations, this study proposes the Improved LSTMs Ensemble with Stacking (ILES) framework for winter RST prediction. Rather than relying on a single recurrent architecture, ILES integrates two base learners with partially complementary predictive characteristics—KNN-LSTM for similarity-driven local pattern retrieval and BiLSTM-MHA for long-range temporal dependency extraction—within a stacking ensemble, where a Bayesian Ridge Regression meta-learner fuses the base learner outputs to yield probabilistic forecasts with closed-form posterior predictive uncertainty at the ensemble combination layer.~~ILES integrates two architecturally complementary base learners—KNN-LSTM for similarity-driven local pattern retrieval and BiLSTM-MHA for long-range temporal dependency extraction—within a leakage-free stacking ensemble, where a Bayesian Ridge Regression meta-learner fuses the base learner outputs to yield probabilistic forecasts with closed-form uncertainty quantification.~~ Three input configurations—a station-only baseline, ERA5-Land reanalysis augmentation, and physics-motivated feature engineering—are systematically compared to determine the most effective and operationally practical input strategy. In addition, SHAP-based attribution analysis is conducted in a stratified manner across temperature regimes and diurnal cycles, providing a more systematic evaluation of whether learned feature importance is consistent with the dominant drivers identified by SEB theory. The framework is trained and evaluated on four consecutive winter seasons of hourly road meteorological observations from station M9393, located in the northwest inland plain of Jiangsu Province, China, with multi-site generalisation assessed at two independent stations M9474 and M9448.

The remainder of this paper is ~~organised~~organized as follows. Section 2 describes the ILES framework, including the KNN-LSTM and BiLSTM-MHA architectures, the stacking ensemble construction, and the Bayesian Ridge Regression meta-learner. Section 3 presents the study site, observational dataset, ERA5-Land reanalysis data, feature engineering procedure, and experimental design. Section 4 reports results encompassing overall predictive performance, input configuration comparisons, ensemble complementarity analysis, and uncertainty quantification. Section 5 ~~summarises~~summarizes the principal conclusions.

2 Methodology

Fig. 1 presents an overview of the ILES framework for winter RST prediction, organised into three sequential stages: data preparation and feature extraction, model training and ensemble construction, and prediction performance evaluation.

In the first stage, historical meteorological observations from station M9393 and ERA5-Land reanalysis variables are subjected to quality control procedures including outlier removal, missing value imputation, and hourly resampling. Input variables are selected on the basis of Spearman rank correlation with winter RST, and the retained variables are standardised using Z-score normalisation parameters estimated exclusively from the training set. A sliding window is applied to construct sequential input samples across three candidate input configurations, enabling the model to capture the multi-hour autocorrelation structure and diurnal periodicity inherent in RST time series.

In the second stage, two structurally distinct LSTM variants serve as base learners within a stacking ensemble. KNN-LSTM augments temporal sequence modeling with similarity-based feature retrieval from historical analogues, enabling the model to leverage locally recurrent meteorological patterns. BiLSTM-MHA employs bidirectional recurrent processing combined with multi-head self-attention to extract long-range temporal dependencies across the full input window. Base learner predictions for the meta-learner training matrix are generated through three-fold temporal cross-validation aligned to complete winter seasons, with each fold corresponding to one complete winter as the validation period. A BRR meta-learner is fitted on the resulting out-of-fold predictions, yielding ~~probabilistic forecasts with closed-form uncertainty quantification as~~ the final ensemble output.

In the third stage, the ILES model is assessed from four analytical perspectives: comprehensive benchmarking against ten models across 1-, 3-, and 6-hour forecasting horizons; comparison of three input variable combinations to evaluate the relative predictive value of station observations, ERA5-Land reanalysis augmentation, and physics-motivated feature engineering; stratified SHAP-based attribution analysis across temperature regimes and diurnal cycles; and multi-site validation at stations with distinct geographical and surface conditions.

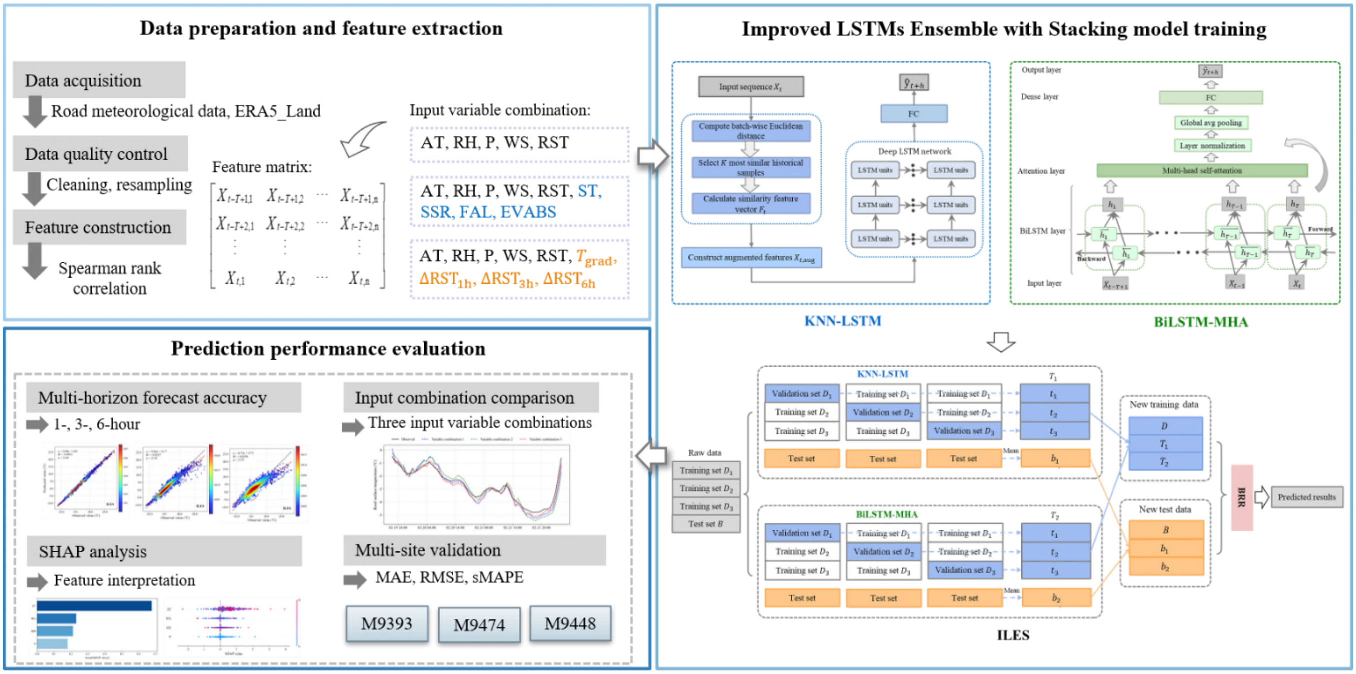


Figure 1. Overview of the ILES model for winter RST prediction.

2.1 KNN-LSTM

The proposed KNN-LSTM model integrates similarity-based feature augmentation with temporal sequence modeling to improve predictive accuracy for wintertime RST forecasting, building upon the framework established by Luo et al. (2019) for traffic flow prediction.

Consider a forecasting task in which the objective is to predict RST at time $t+h$ from a sliding window of historical observations. Let X_t denote the input feature matrix at time t, of dimension $T \times n$, where T is the window size and n is the feature dimension. Given a training set of M historical samples $\{(X_i, y_i)\}_{i=1}^{M}$, where X_i is the i-th input window and y_i is the corresponding target RST, the similarity between a query window X_t and each training sample is measured by Euclidean distance in the flattened feature space. The resulting distance matrix is normalised via min-max scaling to ensure numerical comparability:

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

where ϵ is a small constant to avoid division by zero. The K nearest neighbours $\{X_{(k)}\}_{k=1}^{K}$ are identified as the K reference samples with the smallest normalised distances to X_t. The similarity feature matrix F_t is then constructed as their uniform average:

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

capturing the central tendency of historically analogous meteorological states. The augmented input is formed by concatenating X_t and F_t along the feature dimension:

$$X_{t,\mathrm{aug}} = [X_t, F_t] \ , \tag{4}$$

This augmentation strategy preserves the original temporal structure while incorporating rich contextual information from analogous meteorological conditions in the historical record, and the dimensionality of the input doubles from n to $2n$. The augmented sequence $X_{t,\mathrm{aug}}$ is fed into a three-layer LSTM network (Hochreiter and Schmidhuber, 1997) designed to capture hierarchical temporal dependencies:

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

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

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

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

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

where W_0 and b_0 are the learnable weight matrix and bias term, respectively. A separate instance of this model, with independently optimised parameters $\theta_1, \theta_2, \theta_3$ and W_0, b_0, is trained for each forecasting horizon; no parameters are shared across horizons. A schematic overview of the KNN-LSTM architecture is provided in Fig. 1 and Fig.2.

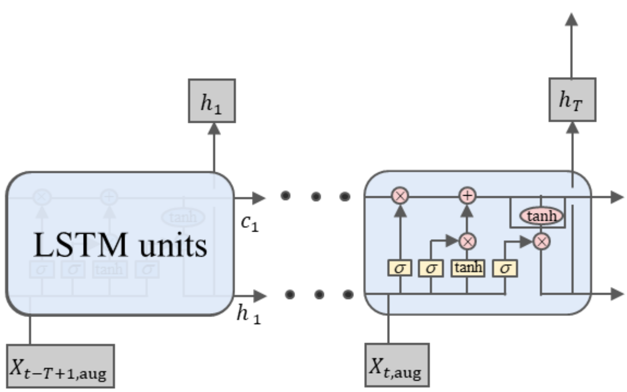


Figure 2. Unrolled computation of a single LSTM layer across the input window, where $\boldsymbol{h_\tau}$ and $\boldsymbol{c_\tau}$ denote the hidden and cell states at time step $\boldsymbol{\tau}$, with $\boldsymbol{\tau} \in \{\mathbf{1}, \dots, \boldsymbol{T}\}$ indexing the steps within the window of length $\boldsymbol{T}$.

2.2 BiLSTM-MHA

The proposed BiLSTM-MHA model integrates a Bidirectional Long Short-Term Memory network with a multi-head self-attention mechanism (Vaswani et al., 2017), augmented by residual connections and layer ~~normalisation~~normalization (Ba et al., 2016) to maintain training stability while capturing complex temporal dependencies.

Let the sliding window size be T and the input sequence at time t be $X = (X_{t-T+1}, X_{t-T+2}, \dots, X_t)^\top$, where $X_t = (X_{t,1}, X_{t,2}, \dots, X_{t,n})$ and n denotes the feature dimension.

The BiLSTM layer captures temporal dependencies in both temporal directions (Schuster and Paliwal, 1997). At each time step t, the forward LSTM processes the sequence from past to future and the backward LSTM processes the same input sequence in the reverse temporal direction, producing a concatenated hidden state $h_t = [\vec{h}_t, \overleftarrow{h}_t]^\top$, where $\vec{h}_t$ and $\overleftarrow{h}_t$ denote the forward and backward hidden states, respectively. The full sequence $H = (h_1, h_2, \dots, h_t)$ is passed to the subsequent attention layer. It should be noted that the backward pass operates exclusively within the historical input window $[t - T + 1, t]$ and accesses no observations beyond time t.

The multi-head self-attention mechanism (Vaswani et al., 2017) transforms H through learned linear projections into multiple representation subspaces in parallel, enabling the model to attend simultaneously to different temporal positions and feature abstractions:

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

where each head is defined as $\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ with $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$. Here W^O denotes the output projection matrix, W_i^Q, W_i^K, W_i^V are the per-head projection matrices, and d_k is the key dimension. In the self-attention formulation adopted here, Q, K, and V are all derived from H.

A residual connection is applied to the attention output to facilitate gradient flow during backpropagation (He et al., 2016), followed by layer normalisation to stabilise the feature distribution:

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

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

Temporal aggregation is then performed by global average pooling across all time steps:

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

Global average pooling reduces parameter count relative to concatenation-based aggregation and confers robustness to sequence length variation (Lin et al., 2013). The final prediction is generated through a fully connected output layer:

$$\hat{y}_{t+h} = W_0\mathrm{c} + b_0 \,, \tag{13}$$

where W_0 represents the weight matrix and b_0 denotes the bias term. As with KNN-LSTM, a separate instance of the BiLSTM-MHA model—including all BiLSTM, attention, and output-layer parameters—is trained independently for each forecasting horizon, with no weight sharing across horizons. A schematic of the BiLSTM-MHA architecture is provided in Fig. 1 and Fig. 3.

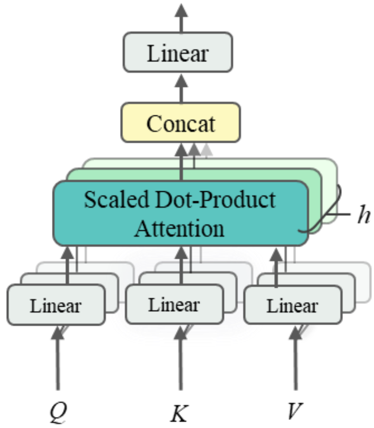


Figure 3. Multi-Head Attention consists of $\boldsymbol{h}$ parallel attention heads, each operating on a distinct learned projection of the input. $\boldsymbol{Q}$, K, and V are all derived from the BiLSTM output.

2.3 Stacking ensemble with out-of-fold cross-validation

Stacking is a hierarchical ensemble learning method that combines predictions from multiple base learners through a meta-learner to achieve superior predictive performance (Wolpert, 1992; Breiman, 1996). Unlike bagging and boosting, which rely on parallel or sequential resampling strategies, stacking employs a two-layer architecture where diverse base models generate intermediate predictions that are subsequently integrated by a meta-model. This framework has demonstrated remarkable success in time series forecasting tasks by leveraging model complementarity (Divina et al., 2018).

A fundamental requirement of stacking ensemble methods is that the meta-learner must be trained on predictions that the base learners generated for samples they did not observe during their own training. Violating this requirement introduces data leakage: base learners that are first trained on the full training set and subsequently used to predict in-sample observations produce fitted values rather than genuine forecasts, causing the meta-learner to learn a combination rule optimized for interpolation rather than generalization (Wolpert, 1992). To satisfy this requirement, we generate base learner predictions for the meta-learner's training matrix through a K-fold temporal cross-validation scheme. Specifically, the model consists of two layers: the first layer comprises N distinct base learners, while the second layer incorporates a meta-learner. The multi-model fusion prediction workflow of Stacking, based on K-fold cross-validation, can be summarized as follows:

(1) Dataset Partitioning.

The original dataset is first partitioned into a training set D and a holdout test set B. The training set D is further subdivided into K disjoint subsets, denoted $D_1, D_2, \dots, D_K$.

(2) Base Learner Training.

For each base learner $L_n (n = 1,2, \dots, N)$, K-fold cross-validation is performed. In the k-th fold, the validation set is D_k and the training set is $D_{(-k)} = D \setminus D_k$. This procedure yields K independently trained instances of each base learner, one per fold.

(3) Construction of the Augmented Training Set.

Each trained instance of base learner L_n generates out-of-fold predictions on its corresponding validation subset D_k. Concatenating these predictions across all K folds ~~produces~~produce a full-length out-of-fold prediction vector T_n for base learner L_n:

$$T_n = \{t_1, t_2, t_3, \dots, t_K\}, \tag{14}$$

The out-of-fold prediction vectors from all N base learners are concatenated with the original training set D to form the augmented training set D', which serves as input for the second-level meta-learner:

$$D' = \{T_1, T_2, T_3, \dots, T_N, D\}, \tag{15}$$

Upon completion of K-fold cross-validation, each base learner L_n produces K prediction vectors on the holdout test set B, denoted $b_n^1, b_n^2, \dots, b_n^K$. These K predictions are averaged to yield a single representative test prediction b_n for base learner L_n. The consolidated test predictions from all N base learners are combined with B to form the augmented test set B':

$$B' = \{b_1, b_2, b_3, \dots, b_N, B\}, \tag{16}$$

(4) Meta-Learner Training.

The augmented dataset D' is used to train the second-level meta-learner, and B' is used to evaluate its generalisation performance. The meta-learner learns an optimal combination strategy over the base learner outputs, yielding a final ensemble prediction that is expected to surpass any individual constituent model.

The number of folds K is determined by the temporal structure of the data. Since the training dataset spans three complete winter seasons, a 3-fold configuration naturally implements a leave-one-season-out protocol in which each fold corresponds to exactly one complete winter as the validation period. This structure mirrors the real-world forecasting scenario of predicting an unseen winter season from prior observations, and fold boundaries are strictly aligned with seasonal transitions with temporal ordering preserved throughout. The complete workflow is illustrated in Fig. 4.

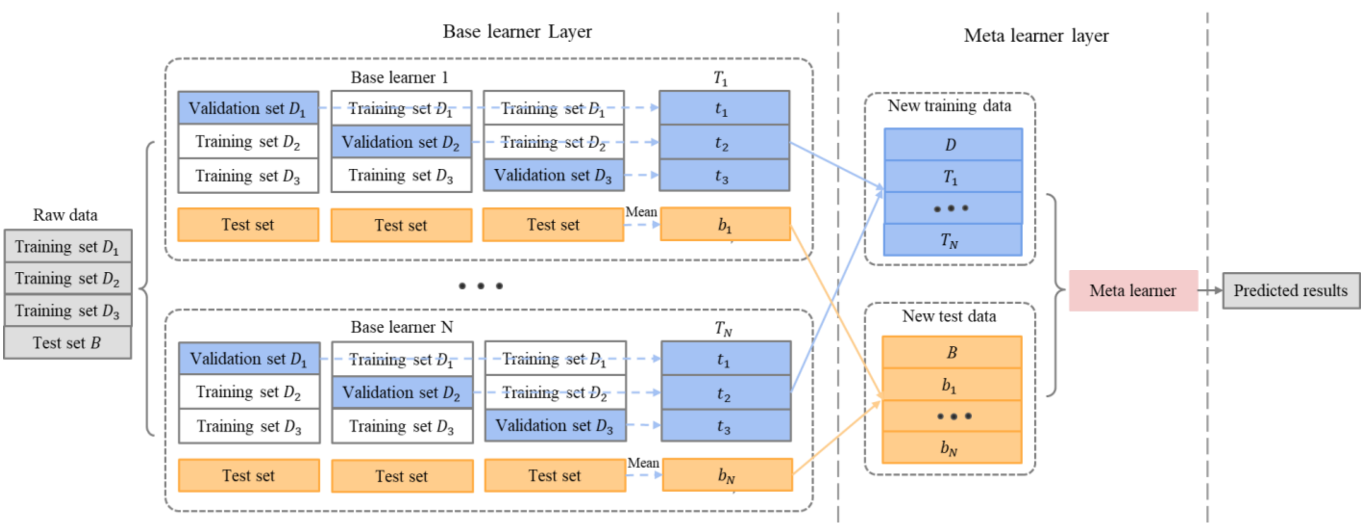


Figure 4. Architecture of the Stacking Ensemble based on Out-of-Fold Cross-Validation.

2.4 Bayesian ridge regression meta-learner

Since the meta-learner operates on base learner predictions rather than raw input features, its primary role is to learn an optimal fusion of these predictions that accounts for their relative strengths and inter-model correlations. We adopt Bayesian Ridge Regression (BRR; MacKay, 1992) as the meta-learner, motivated by three considerations. First, base learner predictions tend to be correlated due to shared input features; the L2 regularization inherent in BRR effectively alleviates multicollinearity (Hoerl and Kennard, 1970). Second, the Bayesian framework automatically determines the optimal regularization strength through Evidence Maximization, precluding overfitting without manual hyperparameter tuning (Bishop, 2006). Third, BRR yields probabilistic outputs that quantify prediction uncertainty (Gelman et al., 2013), a property of practical value for risk-sensitive winter road maintenance decisions.

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

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

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

The posterior distribution over weights is analytically tractable:

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

The precision hyperparameters $\alpha$and $\lambda$are optimised via Type-II Maximum Likelihood (Evidence Maximisation), which automatically balances data fit against regularisation without requiring manual cross-validation.

Beyond point prediction, BRR yields a closed-form posterior predictive distribution for each forecast:

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

where $\sigma_n^2(\mathrm{x}) = \alpha^{-1} + \mathrm{x}^\top \Sigma_n \mathrm{x}$ decomposes total predictive uncertainty into aleatoric noise α^{-1} and parameter uncertainty encoded in the posterior covariance Σ_n. A separate BRR meta-learner is fitted independently for each forecasting horizon, receiving the two-dimensional out-of-fold prediction vector generated by the two base-learner instances trained for that horizon.

2.5 Multi-Horizon Forecasting Strategy

Multi-step forecasting can be approached through four principal strategies, including recursive, direct, hybrid, and multi-input multi-output (MIMO) approaches (Taieb et al., 2012). In the present study, the direct method is adopted: a fully independent instance of the ensemble is trained separately for each forecasting horizon, with no parameter sharing across horizons. This choice avoids the error accumulation inherent to the recursive and hybrid strategies, and avoids the output-layer distribution mismatch that can arise when a single MIMO model must simultaneously minimise losses at horizons with substantially different error magnitudes. The additional training cost of maintaining three independent model instances is computationally manageable given the offline nature of model development for this application.~~Multi-step forecasting encompasses four principal strategies (Taieb et al., 2012): (1) Direct method, which constructs an independent model for each forecasting horizon, incurring high computational cost but avoiding error accumulation. (2) Recursive multi-step method, which iteratively feeds~~

~~predicted outputs as subsequent inputs using a single one-step-ahead model, yet suffers from compounding error propagation over extended horizons. (3) Direct-recursive hybrid method, which integrates both strategies by building horizon-specific models that additionally leverage prior predicted outputs as inputs, though error accumulation remains a concern. (4) Multi-input multi-output (MIMO) method, which simultaneously predicts all target horizons within a single model inference, fundamentally eliminating error accumulation at the cost of increased model complexity.~~

~~In the present study, the MIMO strategy is adopted to achieve superior forecasting accuracy with enhanced computational efficiency. Formally, the MIMO formulation is expressed as:~~

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

~~where $X_1, X_2, \cdots, X_N$ denote the input feature vectors, $Y_{N+1}, Y_{N+2}, \cdots, Y_{N+k}$ represent the multi-step output targets, and $f(\cdot)$ denotes the predictive model.~~

3 Experiments

3.1 Datasets and metrics

3.1.1 Site data

This study focuses on a road segment in the northwest inland plain of Jiangsu Province, China, adjacent to the Longhai Railway overpass. The region belongs to a warm temperate semi-humid monsoon climate zone, with minimum daily temperatures in winter reaching approximately -3°C. The terrain is predominantly flat, which facilitates the horizontal distribution of meteorological variables and ~~minimises~~minimizes orographic effects that would otherwise compromise the spatial representativeness of gridded reanalysis data.

The primary dataset was collected from road meteorological monitoring station M9393, located west of the Longhai Railway overpass at coordinates (34.30°N, 117.04°E), at a temporal resolution of five minutes. The dataset spans four consecutive winter periods: December 2020 to February 2021, December 2021 to February 2022, December 2022 to February 2023, and December 2023 to February 2024. Infrared remote sensing sensors recorded near-surface meteorological variables including visibility (V), air temperature (AT), relative humidity (RH), precipitation (P), wind speed (WS), wind direction (WD), and road surface temperature (RST), ~~totalling~~totaling 102,431 raw observations. After quality control and imputation, the dataset was chronologically partitioned into training and test periods, and the meteorological factors and RST data within each period were independently resampled at hourly intervals, resulting in a refined dataset of 8,664 samples for modeling.

~~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. Following quality control, all variables were resampled to hourly resolution: precipitation was aggregated as the hourly total, and all other variables were computed as hourly means to suppress transient sub-hourly fluctuations. This yielded a final dataset of 8,664 samples for model development.~~

RST exhibits significant diurnal periodicity and lag responses to meteorological forcing, reflecting the complex and nonlinear nature of heat exchange at the pavement–atmosphere interface (Cheng et al., 2021). Fig. 5 illustrates the daily variation of RST during December 2020, showing a consistent diurnal cycle with a fluctuation range of approximately 5°C to 15°C and strong hour-to-hour autocorrelation across successive days. Fig. 6 presents the mean 24-hour diurnal RST profile and its 95% confidence interval computed across the full study period. RST reaches its minimum during the pre-dawn hours, rises gradually through the morning, reaches a mean daily maximum of approximately 6°C around 14:00, and subsequently declines through the evening. The width of the confidence interval reflects the day-to-day variability at each hour arising from meteorological disturbances superimposed on the underlying diurnal cycle. These characteristics motivate the inclusion of historical RST sequence an input feature~~, as the recent RST trajectory encodes information about the current thermal state of the pavement that is not directly available from instantaneous meteorological observations alone~~.

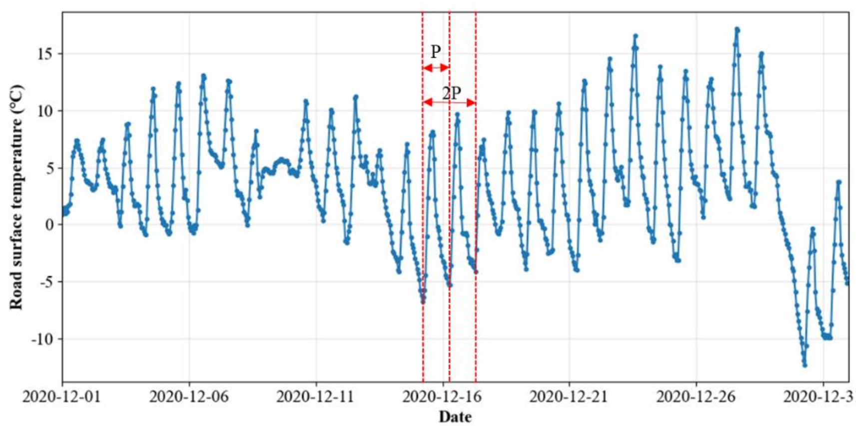


Figure 5. Periodic variation of RST. T represents a period, with time corresponding to one day.

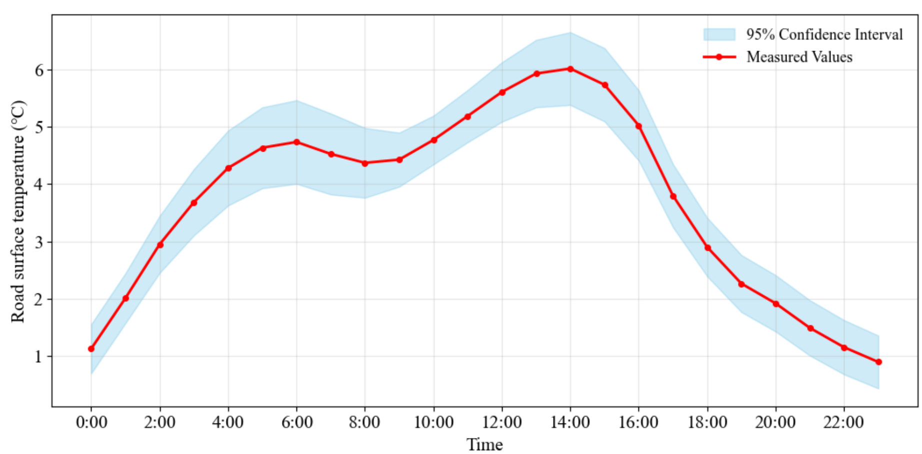


Figure 6. The 24-hour diurnal variation curve of RST. The shaded blue part of the graph is the 95% confidence interval for the RST.

3.1.2 ERA5-Land reanalysis data

ERA5-Land reanalysis data were obtained for the same temporal coverage as the station observations to investigate the potential contribution of physically meaningful auxiliary variables. ERA5-Land is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) through land surface model simulations driven by ERA5 atmospheric forcing, and is publicly available from the Copernicus Climate Change Service (https://cds.climate.copernicus.eu/). The dataset provides hourly estimates at a spatial resolution of 0.1°×0.1° (approximately 9–11 km), representing area-averaged surface conditions rather than point-scale measurements.

Eight candidate variables with established thermodynamic relevance to road surface temperature dynamics were extracted as the initial variable pool: soil temperature (ST), surface net solar radiation (SSR), surface net thermal radiation (STR), surface sensible heat flux (SSHF), surface latent heat flux (SLHF), forecast albedo (FAL), evaporation from bare soil (EVABS), and total evaporation (E). These variables collectively represent the principal radiative, turbulent, and latent heat exchange terms in the surface energy balance, and their physical relevance to RST evolution is established in the literature (Chen et al., 2019; Hermansson, 2004). Variables ultimately retained for model input were determined through the correlation screening procedure described in Section 3.2.

Spatial matching was performed using the nearest-neighbour method, identifying the grid point (34.30°N, 117.00°E) closest to the station. This approach is considered appropriate given the flat terrain of the study region, which ~~minimises~~minimizes the representativeness error associated with nearest-neighbour interpolation (Muñoz-Sabater et al., 2021). All variables were temporally aligned with the station dataset by converting from UTC to China Standard Time (UTC+8). Unit conversions were applied where necessary: SSR, STR, SSHF, and SLHF, provided as hourly accumulated values in $J\cdot m^{-2}$, were divided by 3600 to obtain mean flux densities in $W\cdot m^{-2}$; ST was converted from Kelvin to degrees Celsius.

Table 2. Summary of the dataset.

Feature	Mean	Std	Min	Max	Unit
Visibility	7942.28	3355.89	56.00	30000.00	m
Air temperature	2.08	5.29	-15.15	23.79	°C
Relative humidity	55.97	24.14	1.76	100.00	%
Precipitation	0.01	0.12	0.00	4.60	mm
Wind speed	1.89	1.19	0.30	8.72	$m\cdot s^{-1}$
Wind direction	175.41	92.87	0.03	359.95	°
Road surface temperature	3.72	5.59	-12.99	26.52	°C
Soil temperature	3.83	3.92	-2.60	20.88	°C
Surface net solar radiation	99.42	164.05	0.00	652.08	$W\cdot m^{-2}$

Surface net thermal radiation	68.53	348.85	0.00	2601.48	W·m⁻²
Surface sensible heat flux	30.95	105.62	0.00	1671.75	W·m⁻²
Surface latent heat flux	20.65	109.59	0.00	1522.77	W·m⁻²
Forecast albedo	0.17	0.05	0.15	0.59	-
Evaporation from bare soil	-0.42	0.32	-1.70	0.00	mm
Total evaporation	-0.56	0.37	-2.25	0.00	mm

3.1.3 Evaluation index

In this study, multiple evaluation metrics are employed to comprehensively evaluate the predictive performance of the models, including the mean absolute error (MAE), root mean squared error (RMSE), symmetric mean absolute percentage error (sMAPE), and coefficient of determination (R^2).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \ ,$$ (~~22~~21)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \ ,$$ (~~23~~22)

$$\mathrm{sMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i| + \varepsilon} \times 100\% \ ,$$ (~~24~~23)

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ ,$$ (~~25~~24)

where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the observed values, and n is the total number of samples. and ε is a small constant ($\varepsilon = 10^{-8}$) introduced to prevent division by zero when both the observed and predicted values simultaneously approach zero.

3.2 Feature extraction

To select a parsimonious and informative input feature set, Spearman rank correlation coefficients (Spearman, 1961) were computed between each candidate variable and winter RST. This nonparametric measure is used in preference to Pearson correlation given the potential for nonlinear monotonic relationships between meteorological variables and RST. The coefficient is calculated as:

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2-1)} \ ,$$ (~~26~~25)

where d_i is the rank difference between paired observations, and n is the number of samples.

As illustrated in Fig. 7, among the station-observed variables, air temperature exhibits the strongest positive correlation with RST ($\rho = 0.93$), consistent with its recognised role as the primary meteorological driver of near-surface pavement temperature

(Chen et al., 2019). Wind speed and precipitation display moderate positive correlations of 0.29 and 0.27, respectively, while relative humidity shows a moderate negative correlation ($\rho = -0.25$), consistent with the known cooling effect of high-humidity conditions at the pavement surface (Gui et al., 2007). Visibility and wind direction exhibit correlations below 0.20 and are excluded from the input feature set. 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 is consistent across both conventional and circular-adapted metrics, and the exclusion of wind direction from the final feature set is supported regardless of the association measure applied. The RST sequence is retained as an input variable because, as a directly measured surface quantity, it reflects the integrated effect of prior meteorological forcing at this monitored cross-section and provides information that is complementary to the instantaneous meteorological observations.~~it reflects the integrated effect of prior meteorological forcing and provides information about the current thermal state of the pavement that is complementary to the instantaneous meteorological observations.~~ Based on these findings, the input feature set for the primary analytical perspectives comprises five station-observed variables: AT, RH, P, WS, and historical RST.

For the analytical perspective that investigates the added predictive value of ERA5-Land reanalysis variables (Sect. 4.2, Variable combination 2), the correlation analysis is extended to the full candidate variable pool including both station observations and ERA5-Land fields. Among the ERA5-Land variables, soil temperature shows the strongest positive correlation with RST ($\rho = 0.49$), which may reflect its role as an integrative indicator of subsurface and near-surface thermal conditions. Surface net solar radiation and forecast albedo exhibit correlations of 0.24 and -0.20, respectively, consistent with the known sensitivity of pavement temperature to radiative forcing (Qin et al., 2022). Evaporation from bare soil shows a moderate negative correlation ($\rho = -0.23$), consistent with the cooling effect associated with surface moisture conditions. The remaining ERA5-Land variables exhibit correlations below 0.20 with RST and are excluded. Accordingly, the augmented input set for Variable combination 2 in Sect. 4.2 comprises nine variables: AT, RH, P, WS, ST, SSR, FAL, EVABS, and historical RST.

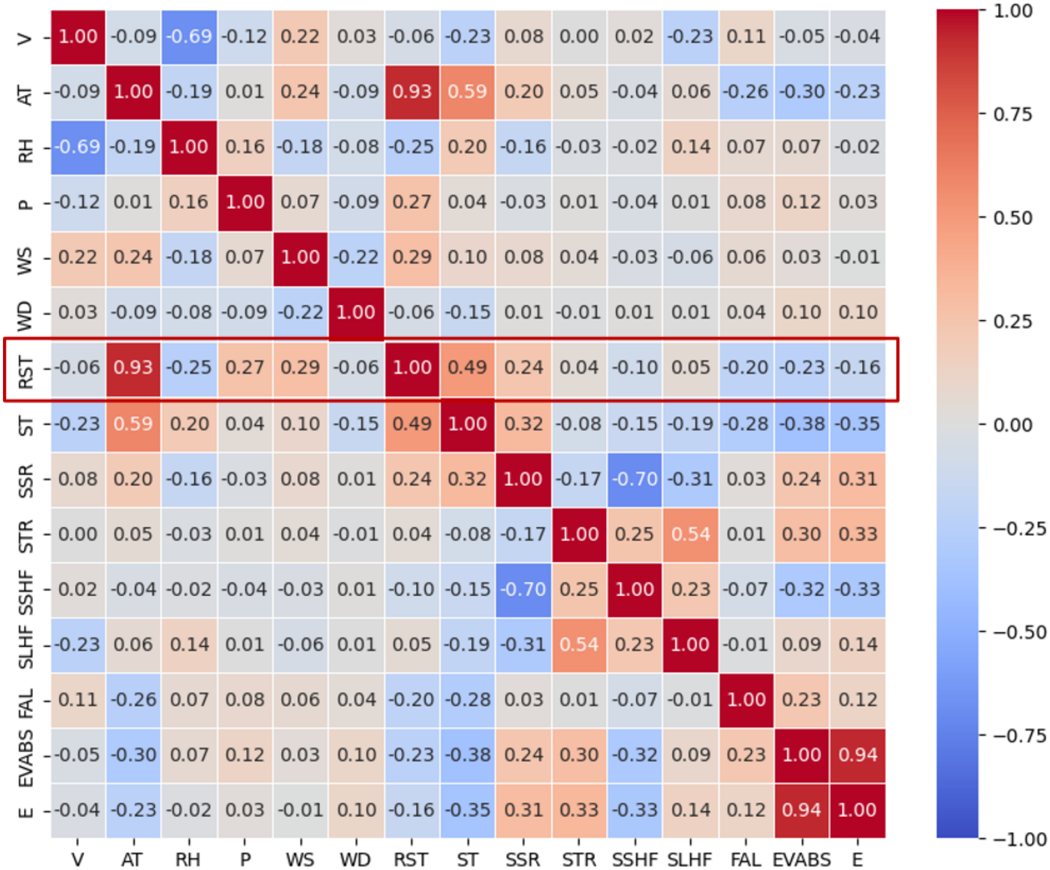


Figure 7. Spearman correlation coefficient plot between RST and various meteorological factors and ERA5_Land physical factors.

3.3 Optimization

3.3.1 Sliding window size

The sliding window method is employed to construct input variables and prediction targets for the model, and the selection of the window size directly influences the ability of the model to capture temperature periodicity and lag effects of meteorological variables (Wu et al., 2020). To identify the optimal window size, we conducted a comparative analysis of the improved base models under varying window configurations (Table 3). This analysis revealed that a 24-hour sliding window offers an effective balance between predictive performance and computational efficiency, while also matching the timescale of the diurnal RST cycle, enabling the model to learn the recurring day–night thermal rhythm directly from the historical RST and meteorological sequence at this site~~while also aligning with the diurnal solar cycle, enabling the model to learn the day–night thermal rhythm directly from the historical RST and meteorological sequence~~ (Zhang et al., 2022). Consequently, the 24-hour window was selected as the standardized configuration for all subsequent modeling experiments.

Table 3. Impact of sliding window size on errors.

Window size	KNN-LSTM			BiLSTM-MHA		
	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)
6-hour	0.507	0.665	27.577	0.487	0.655	25.871
12-hour	0.517	0.669	28.550	0.431	0.617	22.000
24-hour	0.389	0.536	21.348	0.393	0.545	20.783
48-hour	0.414	0.575	22.357	0.662	0.938	30.120
72-hour	0.399	0.540	22.139	0.415	0.569	22.772

3.3.2 Training

The dataset was divided into training and testing subsets. The training subset comprised data from December 1, 2020 to February 28, 2023, and the testing subset comprised data from December 1, 2023 to February 29, 2024. 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:

$$X_{scaled} = \frac{x_{train} - \mu_{train}}{\sigma_{train}},$$ (~~27~~26)

where μ_{train} and σ_{train} are the mean and standard deviation computed from the training set only. Model outputs were inverse-transformed to degrees Celsius prior to evaluation. Accordingly, all reported MAE and RMSE values are in °C.

To assess the climatological representativeness of the test period, the mean air temperature and RST during the test winter were compared against the three-winter training period mean. The test-period mean air temperature (-1.29°C) and mean RST (0.49°C) deviate by less than 1.0°C from the corresponding training-period means (-0.31°C and 1.40°C, respectively), and the proportions of sub-zero RST hours are comparable (45.74% and 47.53%). These statistics confirm that the test winter is climatologically representative of the study period and that the evaluation is not materially affected by anomalous thermal conditions.

In the KNN-LSTM model, the hyperparameter K specifies the number of nearest-neighbour historical sequences used to construct the similarity feature vector F_t, and directly governs the trade-off between local pattern specificity and representational stability. A value of K that is too small risks overfitting to unrepresentative ~~neighbors~~neighbours, while an excessively large K may dilute the local similarity signal by averaging over dissimilar historical states (Li et al., 2021). To determine the optimal K, a grid search over $K \in \{3, 5, 7, 9, 11, 13, 15, 17, 19\}$ was conducted by evaluating MAE and RMSE on the validation set. As shown in Fig. 8, both metrics decrease monotonically with increasing K, reaching their minimum at $K = 15$, beyond which performance stabilizes. Accordingly, $K = 15$ is adopted as the fixed hyperparameter for all subsequent experiments. It should be noted that this search is a one-time offline procedure performed during model development; during inference, the KNN-LSTM model executes a single forward pass using the fixed $K = 15$, incurring no additional computational overhead.

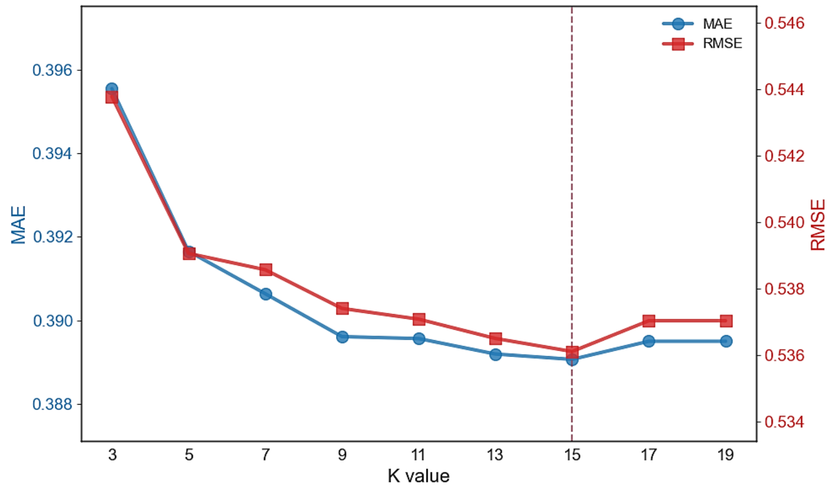


Figure 8. The effect of K value on MAE and RMSE.

The KNN-LSTM model employs a three-layer LSTM architecture with 64, 32, and 16 hidden units in successive layers, with tanh activation in the first layer and ReLU in deeper layers to mitigate the vanishing gradient problem (Glorot et al., 2011). The BiLSTM-MHA model uses a bidirectional LSTM with 128 units per direction, followed by a multi-head attention mechanism with 4 heads and key dimension 64, layer normalization with residual connection, global average pooling, and a single-unit dense output layer. Both models are trained with the Adam optimizer (Kingma and Ba, 2014) at a learning rate of 0.001, mean squared error loss, batch size 32, a maximum of 100 epochs, and early stopping with patience 20 to prevent overfitting. L2 regularization is applied to all LSTM layers to further improve generalization. This architecture and training configuration is shared across the three forecasting horizons; for each horizon, a separate model instance with this identical configuration is trained independently, yielding horizon-specific weights with no parameter sharing across horizons.~~The KNN-LSTM model employs a three-layer LSTM architecture with 64, 32, and 16 hidden units in successive layers, with tanh activation in the first layer and ReLU in deeper layers to mitigate the vanishing gradient problem (Glorot et al., 2011). The BiLSTM-MHA model uses a bidirectional LSTM with 128 units per direction, followed by a multi-head attention mechanism with 4 heads and key dimension 64, layer normalization with residual connection, global average pooling, and a single-unit dense output layer. Both models are trained with the Adam optimizer (Kingma and Ba, 2014) at a learning rate of 0.001, mean squared error loss, batch size 32, a maximum of 100 epochs, and early stopping with patience 20 to prevent overfitting. L2 regularization is applied to all LSTM layers to further improve generalization.~~

3.4 Complementarity and ablation analysis

The effectiveness of a stacking ensemble depends critically on base learner diversity: models that commit errors in different input regions enable the meta-learner to achieve superior performance beyond any individual component. Computational constraints limit the number of base learners, as each requires independent out-of-fold training across all folds. KNN-LSTM and BiLSTM-MHA are selected for their architectural complementarity with respect to winter RST spatiotemporal structure. KNN-LSTM integrates similarity-based feature augmentation with sequential ~~modeling~~modelling to identify local pattern recurrence through retrieval of analogous historical meteorological states, making it effective at capturing regime-specific ~~behaviors~~behaviours tied to locally recurring weather conditions. BiLSTM-MHA employs bidirectional recurrent processing with multi-head self-attention to dynamically weight temporal features across the full input window, demonstrating superior capability in extracting global temporal dependencies and long-range patterns during thermally complex periods involving sustained trends or abrupt transitions.

To empirically validate this complementarity prior to ensemble construction, we conduct a complementarity analysis and an ablation study, both at the 1-hour forecasting horizon where the ensemble signal is cleanest. Results are presented in Fig. 9 and Table 4. The pearson correlation coefficient between the residual series of the two base learners is $r = 0.889$, indicating substantial but imperfect co-variation in prediction errors. However, a high global correlation does not preclude conditional complementarity across distinct operating regimes (Kuncheva and Whitaker, 2003). As shown in Fig. 9b, the two models exhibit asymmetric error patterns across temperature regimes: under sub-zero conditions, BiLSTM-MHA achieves a lower MAE, whereas under above-zero conditions KNN-LSTM is more accurate. This regime-dependent asymmetry is noteworthy because sub-zero and above-zero conditions are typically associated with distinct meteorological forcing characteristics — sustained radiative cooling under clear skies versus more variable radiation-driven heating, respectively (Hermansson, 2004; Qin et al., 2022) — suggesting that the two architectures may be exploiting different statistical regularities in the RST signal. Fig. 9c shows comparable daytime and nighttime performance at the aggregate level, indicating that temperature-regime complementarity, rather than the diurnal cycle per se, is the dominant source of predictive diversity between the two base learners. Fig. 9d further confirms that neither model achieves consistent sample-level dominance, demonstrating sustained bidirectional predictive diversity that the meta-learner can systematically exploit.

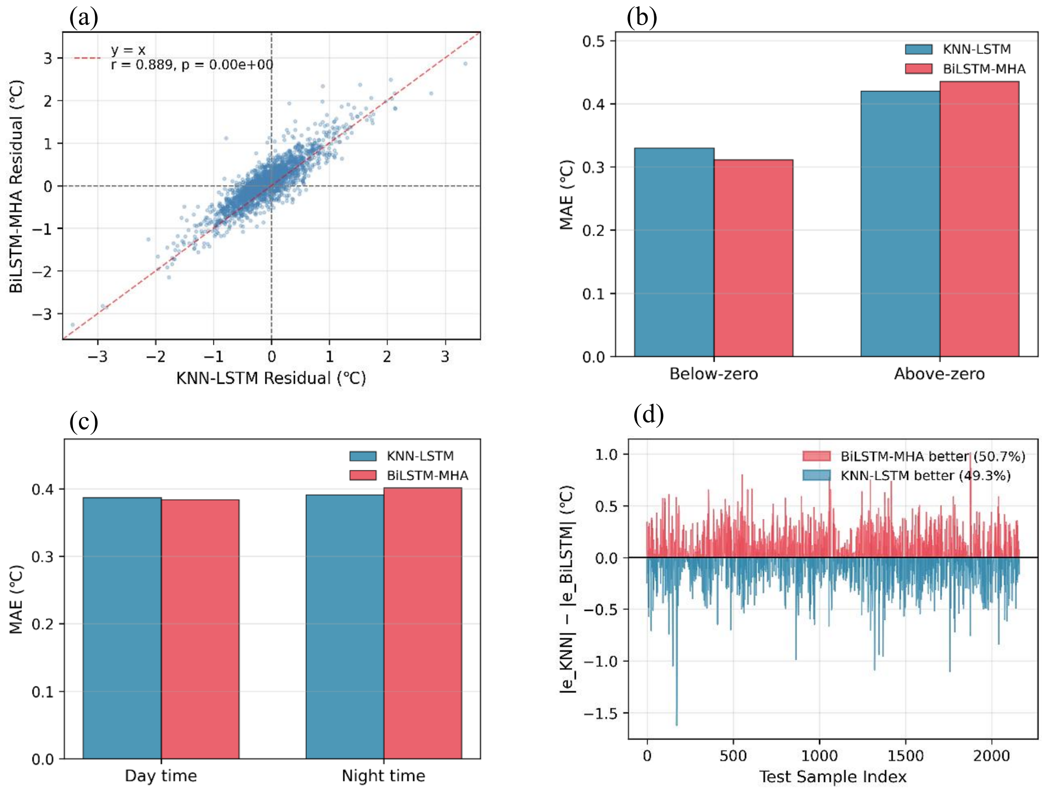


Figure 9. Complementarity analysis of base learners: (a) residual correlation analysis, (b) temperature interval error decomposition, (c) day and night time error decomposition, and (d) time series advantage distribution.

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 reduces MAE by 0.035°C (8.20%), whereas introducing KNN-based similarity augmentation alone 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 partially complementary rather than redundant or synergistic contributions~~complementary rather than redundant contributions~~.

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

3.5 Prediction performance evaluation

To systematically evaluate the proposed ILES framework and address the four objectives stated in Sect. 1, four analytical perspectives are designed as follows.

(1) Comprehensive performance benchmarking.

ILES is evaluated against ten models spanning three categories: a naive persistence baseline, nonlinear regression, traditional machine learning methods (Random Forest, XGBoost), standard deep learning architectures (GRU, LSTM, BiLSTM, CNN-LSTM), and the two proposed base learners (KNN-LSTM, BiLSTM-MHA). All models receive identical station-only inputs and are evaluated at 1-, 3-, and 6-hour forecasting horizons using MAE, RMSE, sMAPE, and R^2.

(2) Input configuration comparison.

Three input variable combinations are evaluated at the 1-, 3-, and 6-hour forecasting horizons to assess the relative predictive value of different information sources. Variable combination 1 comprises the five directly observed meteorological variables: AT, RH, P, WS, and lagged RST. Variable combination 2 extends Variable combination 1 with four ERA5-Land reanalysis variables — ST, SSR, FAL, and EVABS — selected on the basis of Spearman rank correlation with winter RST. Variable combination 3 extends Variable combination 1 with four derived features constructed exclusively from existing station observations T_{grad}, and RST temporal tendency features at 1-, 3-, and 6-hour lags ($\Delta\mathrm{RST}_{1\mathrm{h}}$, $\Delta\mathrm{RST}_{3\mathrm{h}}$, $\Delta\mathrm{RST}_{6\mathrm{h}}$).

(3) Feature importance and model interpretability.

SHAP analysis is applied to ILES under the station-only configuration to quantify the marginal contribution of each input feature to individual predictions. Attribution values are examined stratified by temperature regime and across the diurnal cycle. The resulting feature importance patterns are qualitatively compared with expectations derived from established meteorological understanding of near-surface heat exchange processes, serving as a plausibility check on the model's learned input–output ~~behavior~~behaviour.

(4) Multi-site validation.

The cross-site applicability of the ILES framework is assessed at two independent stations, M9474 and M9448, using station-only inputs at the 1-hour horizon. At each station, the framework is retrained independently on site-specific observations following the same protocol as the primary site. Performance rankings across five deep learning models are compared at all three stations to determine whether the predictive advantages of the ensemble are site-specific or structurally consistent across different road environments within the temperate monsoon climate zone. Uncertainty quantification provided by the BRR meta-learner is additionally assessed through a reliability diagram on the held-out test set.

4 Results

4.1 Multi-horizon prediction performance

Table 5 presents the prediction performance of eleven models across 1-, 3-, and 6-hour forecasting horizons. The model hierarchy spans a naive persistence forecasting (assuming RST at time $t + h$ equals RST at time t), a nonlinear regression baseline (NR), traditional machine learning methods (RF, Darghiasi et al., 2025; XGBoost, Kebede et al., 2024), standard deep learning architectures (GRU, LSTM, BiLSTM, CNN-LSTM, Tabrizi et al., 2021), the two proposed base learners (KNN-LSTM, BiLSTM-MHA), and the proposed ensemble ILES.

At the 1-hour horizon, persistence yields an MAE of 0.897°C, RMSE of 1.295°C, and sMAPE of 35.829%, establishing a lower bound of predictive skill that any learned model should substantially exceed. ILES reduces MAE by 58.42% relative to persistence and by 52.96% relative to NR, indicating that the learned temporal representations capture substantially more information than either the most recent observation or a simple parametric reference model. Relative to the traditional machine learning baselines, RF and XGBoost, ILES reduces MAE by 28.82% and 32.67% respectively. Relative to the four standard deep learning architectures, GRU, LSTM, BiLSTM, and CNN-LSTM, ILES reduces MAE by 8.13% to 17.66%, with the smallest margin against CNN-LSTM, the strongest of the four. Relative to its own base learners, ILES reduces MAE by 4.11% over KNN-LSTM and 5.09% over BiLSTM-MHA.

At the 3-hour horizon, the performance gap between persistence and all learned models widens further, consistent with the increasing value of temporal modeling at longer lead times. ILES reduces MAE by 49.22% relative to persistence and 41.35% relative to NR. Against the traditional machine learning and standard deep learning baselines collectively, RF, XGBoost, GRU, LSTM, BiLSTM, and CNN-LSTM, ILES reduces MAE by 5.72% to 12.73%. Relative to its own base learners, the margin narrows to 1.32% over KNN-LSTM, the strongest individual model at this horizon, and 10.39% over BiLSTM-MHA.

At the 6-hour horizon, persistence skill degrades substantially, underscoring the practical necessity of learned models at extended lead times. ILES reduces MAE by 51.11% relative to persistence and 36.00% relative to NR. Among the six traditional machine learning and deep learning baselines, ILES achieves lower MAE than five, with reductions ranging from 4.57% to 10.90%, while GRU attains a marginally lower MAE than ILES by 5.03%. ILES nonetheless maintains the lowest RMSE among all eleven models at this horizon. Relative to its own base learners, ILES reduces MAE by 2.86% over KNN-LSTM and 3.92% over BiLSTM-MHA. sMAPE values increase substantially across all models at this horizon, reflecting the prevalence of near-zero RST values in winter conditions, which inflates this metric independent of absolute forecast accuracy; MAE and RMSE remain the more informative indicators at extended horizons.

Across all three horizons, ILES achieves the lowest MAE and RMSE among the eleven evaluated models, with the single exception of GRU's marginally lower MAE at 6 hours. sMAPE rankings are less consistent: KNN-LSTM attains a marginally lower sMAPE at 3 hours, and GRU, XGBoost, and RF attain lower sMAPE at 6 hours. The magnitude of improvement varies predictably with baseline category and forecasting horizon: gains over persistence and NR remain large at every horizon, ranging from 36% to 58%, while gains over the established machine learning and deep learning baselines narrow from roughly 8–33% at 1 hour to 5–13% at 3 hours and 5–11% at 6 hours, and the margin over the two base learners is smaller still, between 1% and 10% across horizons. This gradient is consistent with a model that captures genuine temporal structure beyond what persistence or simple regression can represent, while offering a more modest but directionally consistent refinement over architecturally related deep learning baselines and its own constituent base learners.~~At the 1-hour horizon, the persistence focasting achieves an MAE of 0.897°C, RMSE of 1.295°C, and sMAPE of 35.829%, establishing the lower bound of predictive skill that any meaningful model must substantially exceed. ILES achieves the lowest errors across all metrics, reducing MAE by 58.42% relative to persistence, confirming that the learned temporal representations provide information far beyond what is encoded in the most recent observation alone. Relative to NR, ILES reduces MAE by 52.96%, confirming that the added complexity of the stacking ensemble is well justified over simple parametric reference models. Compared to the best-performing traditional machine learning model RF, ILES achieves a 28.82% MAE reduction, demonstrating the superiority of deep temporal modeling over feature-agnostic ensemble methods. Among deep learning baselines, CNN-LSTM performs most competitively, yet ILES still achieves an 8.13% improvement. Relative to its own base learners, ILES outperforms KNN-LSTM and BiLSTM-MHA by 4.11% and 5.09% in MAE respectively, confirming that the meta-learner successfully exploits their complementary strengths.~~

~~At the 3-hour horizon, the widening performance gap between persistence and all learned models reflects the increasing predictability gain that temporal modeling delivers as forecast lead time grows. ILES achieves the lowest MAE and RMSE, corresponding to a 49.22% MAE reduction over persistence and confirming that the ensemble continues to extract substantial predictive signal at medium-range horizons. Compared to KNN-LSTM, the strongest individual model at this horizon, ILES reduces MAE and RMSE by 1.32% and 4.72% respectively. Traditional machine learning methods and standard deep learning models show substantially higher errors, with ILES achieving 10.37–12.73% lower MAE relative to this group.~~

~~At the 6-hour horizon, persistence performance degrades substantially, underscoring the practical indispensability of learned models for operational forecasting beyond the immediate near-term. ILES maintains the best overall performance, achieving a 51.09% MAE reduction over persistence and MAE reductions of 2.86% and 3.92% over KNN-LSTM and BiLSTM-MHA respectively. The gap relative to traditional baselines widens further at this horizon, with persistence, NR, RF, and LSTM exhibiting MAE values 104.56%, 56.27%, 7.15%, and 12.23% higher than ILES respectively. sMAPE values increase substantially across all models at this horizon due to the prevalence of near-zero RST values in winter datasets; MAE and RMSE remain the more informative and operationally relevant indicators at extended horizons.~~

~~In summary, ILES demonstrates the most consistent and balanced performance profile across all horizons and metrics. Its advantage over simple baselines validates the necessity of the proposed architectural complexity; its advantage over established machine learning methods confirms the value of deep temporal modeling; and its consistent improvement over its own base~~

~~learners across all three horizons establishes that the stacking ensemble provides genuine and durable predictive gains beyond what either component achieves independently.~~

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

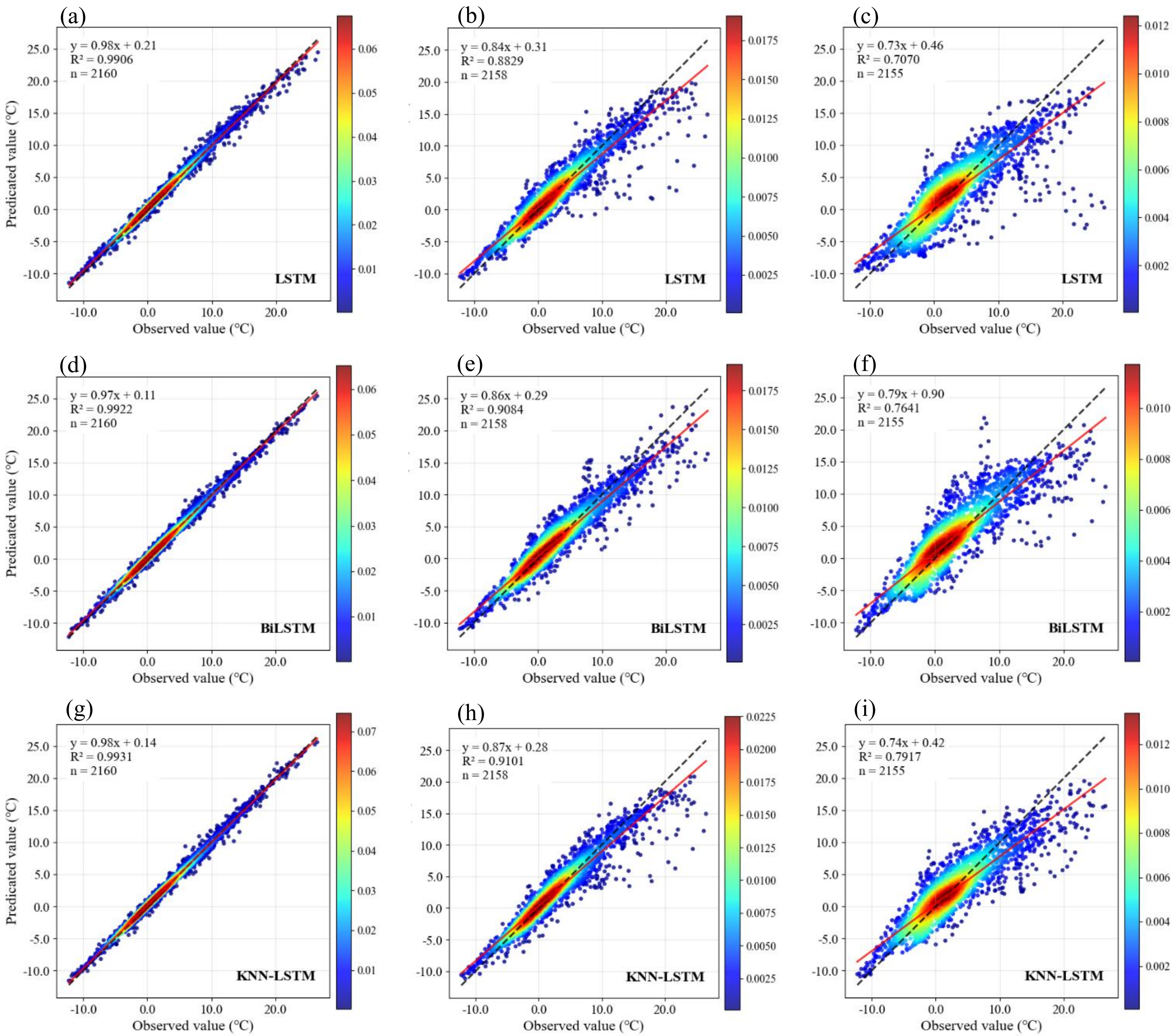
Model	1-hour			3-hour			6-hour		
	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)	MAE (°C)	RMSE (°C)	sMAPE (%)
Persistence	0.897	1.295	35.829	2.497	3.502	72.193	4.312	5.788	101.730
NR	0.793	1.628	34.895	2.162	4.392	62.240	3.294	7.023	76.720
RF	0.524	0.779	24.777	1.415	1.975	51.337	2.259	3.281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

Fig. 10 presents density scatter plots of predicted versus observed RST for five models across three forecasting horizons, enabling both cross-horizon and cross-model comparison of goodness-of-fit and error structure.

Across all models, predictive accuracy degrades systematically with forecast lead time. At the 1-hour horizon, scatter points cluster tightly along the $y = x$ line with high-density regions concentrated near the origin, reflecting strong model fit. At 3 and 6 hours, progressive dispersion is observed: high-density regions contract, outliers increase in frequency, and regression slopes deviate increasingly from unity, indicating that all models struggle to capture rapid thermal transitions at extended horizons.

Across models at the 1-hour horizon, LSTM achieves $R^2 = 0.9906$ with a regression slope of 0.98, while BiLSTM, KNN-LSTM, and BiLSTM-MHA show progressive improvement consistent with their architectural advances. ILES attains the highest goodness-of-fit, with scatter points most tightly concentrated along the ideal line. At the 3-hour horizon, the performance hierarchy is preserved: ILES achieves $R^2 = 0.9227$, outperforming LSTM, BiLSTM, KNN-LSTM, and BiLSTM-MHA. At the 6-hour horizon, ILES maintains $R^2 = 0.8259$, compared to 0.7070 for LSTM, whose scatter plot exhibits the broadest dispersion and most prominent systematic underestimation at high RST values.

A consistent pattern across all models and horizons is that prediction errors increase with RST magnitude, as evidenced by the progressive transition from high- to low-density regions at elevated temperatures. This effect is most pronounced at the 6-hour horizon, where outlier frequency increases sharply above 10°C across all architectures, likely reflecting the greater complexity and variability of RST dynamics under daytime heating conditions, which are not directly represented in the model inputs~~. This behavior likely reflects the greater complexity and stochasticity of RST dynamics under daytime solar heating conditions, which are not directly represented in the model inputs~~ (Qin et al., 2022). Notably, ILES exhibits the most restrained error growth with temperature, suggesting that ensemble integration partially compensates for this systematic limitation.



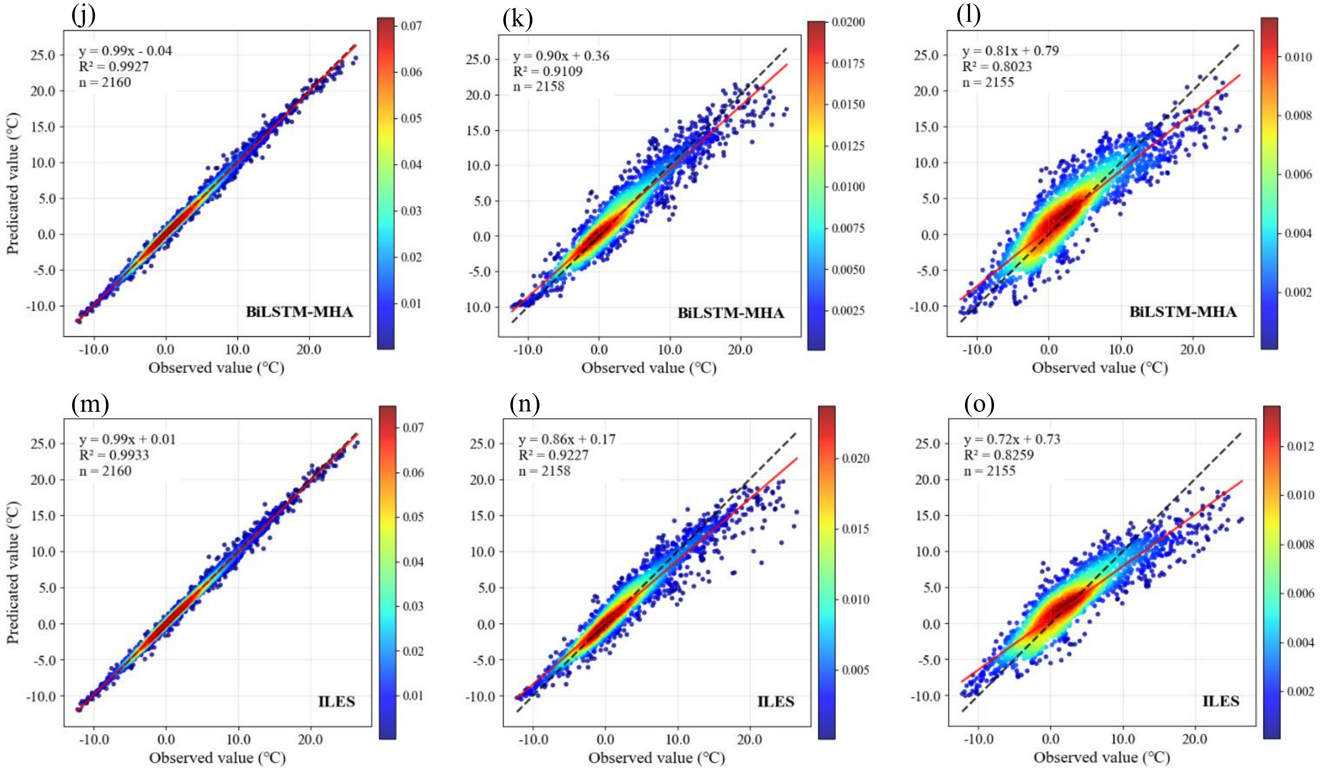


Figure 10. Density scatter plots of predicted versus observed winter RST across 1-hour (a, d, g, j, m), 3-hour (b, e, h, k, n), and 6-hour (c, f, i, l, o) forecasting intervals. Here, ~~colors~~colours indicate the kernel density estimation (KDE) of point concentration, with red representing high-density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

4.2 Input variable combinations comparison

Three input configurations are evaluated to investigate the relative predictive value of different information sources and to determine whether domain-knowledge-guided feature construction provides a practically deployable alternative to reanalysis augmentation for operational RST forecasting.

(1) Variable combination 1 comprises AT, RH, P, WS, and historical RST, corresponding to the standard observational feature set adopted in Sect. 4.1 and serving as the reference configuration.

(2) Variable combination 2 supplements Variable combination 1 with ERA5-Land-derived variables, including ST, SSR, FAL, and EVABS, representing an attempt to recover physically meaningful forcing information through grid-scale reanalysis surrogates.

(~~3)Variable~~3) Variable combination 3 supplements Variable combination 1 with four physically motivated features derived directly from existing station observations. The surface–air temperature difference, defined as $T_{\mathrm{grad}} = T_{\mathrm{RST}} - T_{\mathrm{AT}}$, encodes the sign and magnitude of the near-surface thermal contrast at the pavement–atmosphere interface. This quantity serves as a first-order indicator of the direction of turbulent sensible heat exchange: positive values correspond to daytime conditions under which the road surface is warmer than the overlying air, while negative values are characteristic of nocturnal radiative cooling and are commonly associated with elevated icing risk (Hermansson, 2004). The RST temporal tendency features $\Delta\mathrm{RST}_{\tau}$ for $\tau \in \{1, 3, 6\}$ ~~—h~~ hours quantify the rate of change of road surface temperature over multiple time scales, providing explicit multi-scale descriptors of pavement thermal dynamics alongside the absolute RST sequence. T_{grad} provides a direct, station-measured indicator of the surface–air thermal contrast, and $\Delta\mathrm{RST}_{\tau}$ provides an empirical, multi-scale descriptor of the rate of RST change; both are physically motivated by the surface energy balance framework introduced in Section 1, without requiring additional sensor infrastructure.~~Together,~~ ~~T_{grad} and $\Delta\mathrm{RST}_{\tau}$~~ ~~approximate, respectively, the surface–air thermal contrast and the rate of heat storage change in the pavement substrate — the two dominant flux terms governing RST evolution identified by the surface energy balance framework in Section 1 — without requiring additional sensor infrastructure beyond standard road meteorological station instrumentation.~~ Each feature combination is subsequently used as the input feature vector X fed into the ILES model.

Performance under sub-zero conditions is assessed on the longest continuous segment of test samples satisfying RST < 0°C, spanning from 16:00 on 19 February 2024 to 03:00 on 22 February 2024 and comprising 60 consecutive hourly samples. This segment-based evaluation is adopted to provide a coherent assessment of model behavior during a sustained cooling event, which constitutes the most operationally critical scenario for road icing risk assessment (Song et al., 2023; CMA, 2018).

As shown in Table 6, Variable combination 3 achieves the lowest MAE and RMSE across all three forecasting horizons. Its sMAPE is also lowest at the 1-hour and 3-hour horizons.~~As shown in Table 6, Variable combination 3 achieves the lowest errors across all metrics and forecasting horizons.~~ At the 1-hour horizon, it attains an MAE of 0.19°C and RMSE of 0.23°C, representing reductions of 28.70% and 30.10% relative to Variable combination 1, and 42.60% and 45.10% relative to Variable combination 2. ~~This advantage is sustained and amplified at extended horizons: a~~At 3 hours, Variable combination 3 reduces MAE by 43.50% and RMSE by 45.50% relative to Variable combination 1, and by 48.00% and 41.20% relative to Variable combination 2. At 6 hours, Variable combination 3 continues to yield the lowest MAE and RMSE, while Variable combination 2 remains the weakest performer across all metrics and horizons.

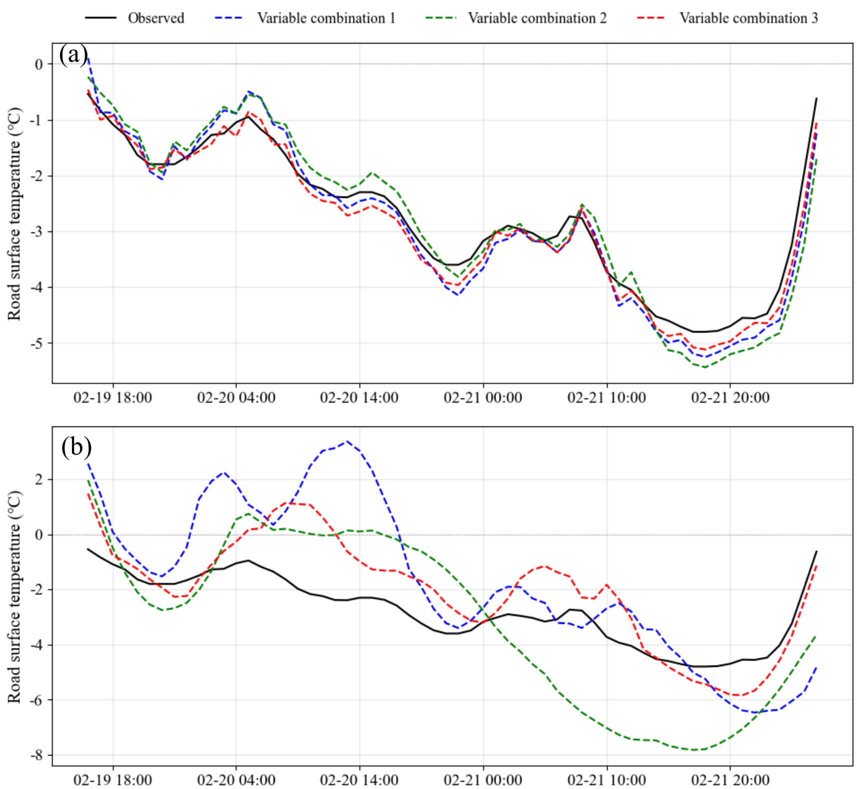
Figure 11 corroborates this hierarchy visually. At the 1-hour horizon, Variable combination 3 tracks the observed cooling descent from 20 to 21 February most faithfully, while combinations 1 and 2 exhibit systematic underestimation of the cooling rate. This behavior is consistent with the role of $\mathrm{T}_{\mathrm{grad}}$ and $\Delta\mathrm{RST}_{\tau}$ in supplying explicit thermal contrast and rate-of-change information during periods of sustained temperature decline, when these signals are most informative. At 3- and 6-hour horizons, all configurations show progressive amplitude attenuation; Variable combination 2 exhibits the most pronounced phase lag and amplitude underestimation, while Variable combination 3 retains the closest agreement with observed thermal minima.

The degraded performance of Variable combination 2 is attributable to two mechanisms. First, ERA5-Land provides area-averaged estimates at approximately 9 to 11 km resolution, whereas RST is governed by point-scale surface conditions; this spatial representativeness mismatch introduces estimation errors that reduce rather than enhance predictive content, a well-documented limitation of gridded reanalysis products in surface applications (Muñoz-Sabater et al., 2021). Second, the historical

RST sequence included in all three configurations already implicitly encodes the net effect of prior meteorological forcing at the pavement surface, rendering the reanalysis-derived surrogates largely redundant while simultaneously increasing input dimensionality (Reichstein et al., 2019). These conclusions are specific to the instrumented, data-rich setting examined here; in station-sparse regions, reanalysis variables may retain supplementary predictive value provided that appropriate bias correction is applied to mitigate the identified scale mismatch.

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



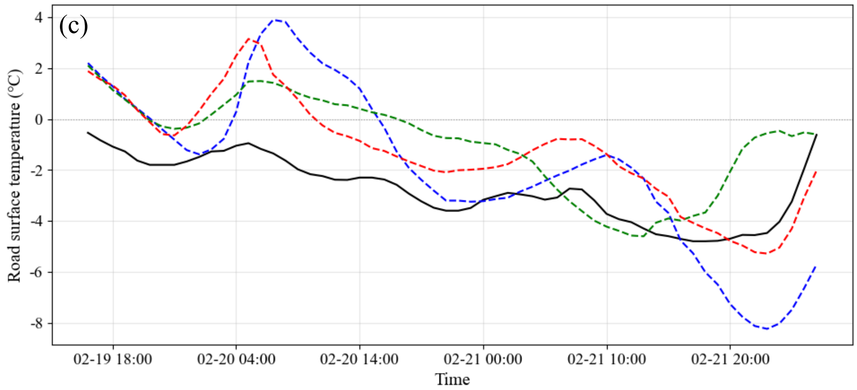


Figure 11. Winter RST prediction of ILES model with three input variable combinations in subzero low temperature period across 1-hour (a), 3-hour (b), and 6-hour (c) forecasting intervals. This period is the longest continuous section of the test sample with RST < 0 °C.

To assess whether the performance hierarchy established under sub-zero conditions reflects a general advantage of Variable combination 3 rather than a regime-specific result, two three-day periods are further examined: a stable clear-sky period (25 to 27 January 2024, ~~characterised~~characterized by mean relative humidity below 50%, zero precipitation, and wind speeds below 2 m/s) and an overcast rainy period (23 to 25 February 2024, ~~characterised~~characterized by continuous precipitation and relative humidity exceeding 85%). These two periods represent contrasting RST variability regimes, spanning periodically dominated and stochastically driven conditions, with direct relevance to operational winter road maintenance (Darghiasi et al., 2023).

Under overcast and rainy conditions (Fig. 12a to c), RST variability becomes markedly more subdued and irregular. Variable combination 2 exhibits the most severe performance degradation across all horizons, with persistent positive bias and substantial loss of observed thermal structure at 6 hours, attributable to the introduction of spatially mismatched reanalysis information under precipitation-dominated conditions. Variable combination 3 consistently maintains the closest correspondence with observations, with its relative advantage most pronounced at 3- and 6-hour.

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.

Taken together, the results across sub-zero, clear-sky, and overcast regimes consistently favour Variable combination 3, confirming that physics-motivated feature engineering provides the most effective ~~and operationally practical~~ input strategy for station-based winter RST prediction at instrumented sites. Accordingly, Variable combination 1 is retained as the standard input configuration for Sections 4.3 and 4.4. Since all ten benchmark models in Section 4.1 were evaluated under this configuration, retaining it as the reference ensures that SHAP-based attribution and cross-site performance rankings remain directly comparable to the established benchmarks without confounding from input differences. It should be noted that in operational settings where physics-motivated feature construction is feasible, Variable combination 3 is recommended as the preferred input strategy given its consistent predictive superiority demonstrated across all forecasting horizons and meteorological regimes.

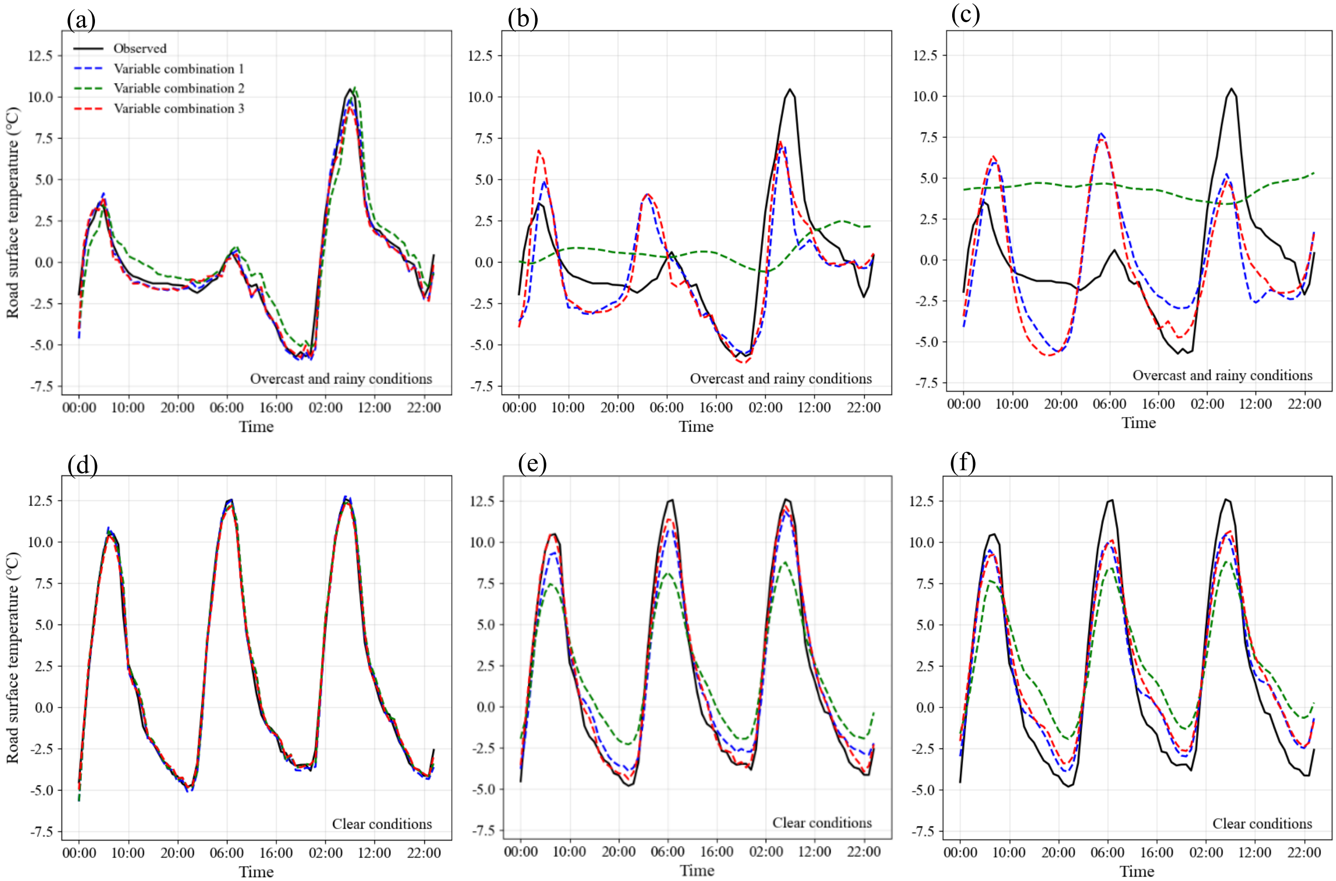


Figure 12. Winter RST prediction of ILES model with three input variables in overcast and rainy and clear synoptic conditions across 1-hour (a, d), 3-hour (b, e), and 6-hour (c, f) forecasting intervals.

4.3 Feature importance and model interpretability

To examine the degree to which the learned input–output relationships of the ILES model are consistent with established meteorological understanding of RST dynamics, SHAP analysis is applied under the station-only input configuration (Variable combination 1). Since the BRR meta-learner operates on base learner predictions rather than raw meteorological features, SHAP values cannot be applied directly to the ensemble output to attribute importance to the original inputs. Instead, KernelSHAP (Lundberg and Lee, 2017) is applied independently to each base learner (KNN-LSTM and BiLSTM-MHA) with respect to the original input feature set, treating each base learner as a black-box function mapping raw features to RST predictions. The resulting SHAP values from the two base learners are aggregated as a weighted average, with weights proportional to the BRR meta-learner coefficients, to produce a single set of feature attributions representative of the full ensemble. This procedure provides a model-agnostic and theoretically principled attribution of ensemble predictions to the original meteorological inputs, grounded in cooperative game theory (Lundberg and Lee, 2017; Joo et al., 2023). Attribution values are examined stratified by temperature regime and across the diurnal cycle, enabling a systematic assessment of whether the learned importance structure is qualitatively consistent with the dominant meteorological drivers identified by surface energy balance theory.

The SHAP analysis presented here is restricted to the four instantaneous meteorological input variables (AT, RH, WS, P). The reported SHAP importances therefore characterize the marginal contributions of the contemporaneous meteorological state, not the full input space. AT is the dominant predictor (Fig. 13a), with a mean absolute SHAP value of 0.696, accounting for 55.4% of total feature importance. RH, WS, and P follow in descending order, contributing 16.9%, 14.3%, and 13.4% respectively. This ranking is consistent with the ~~recognised~~recognized role of AT as the primary driver of RST through near-surface sensible heat exchange, and the secondary modulating roles of RH, WS, and P through evaporative, convective, and latent heat processes (Chen et al., 2019; Gui et al., 2007; Feng and Feng, 2012). The beeswarm plot (Fig. 13b) confirms the expected directionality: high AT values are associated with positive SHAP contributions across the full temperature range, while elevated RH and P values are predominantly associated with negative contributions, consistent with their cooling influence under high-humidity and wet-surface conditions. The narrow spread of P points reflects its episodic character, with predictive influence concentrated in a small fraction of precipitation hours.

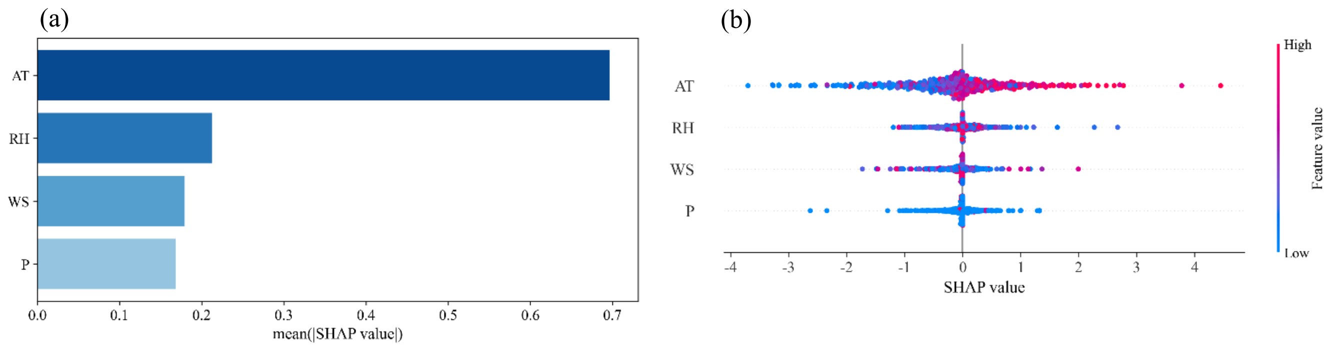


Figure 13. Mean absolute SHAP values (a) and beeswarm plots of SHAP value distributions (b) for the ILES model.

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

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

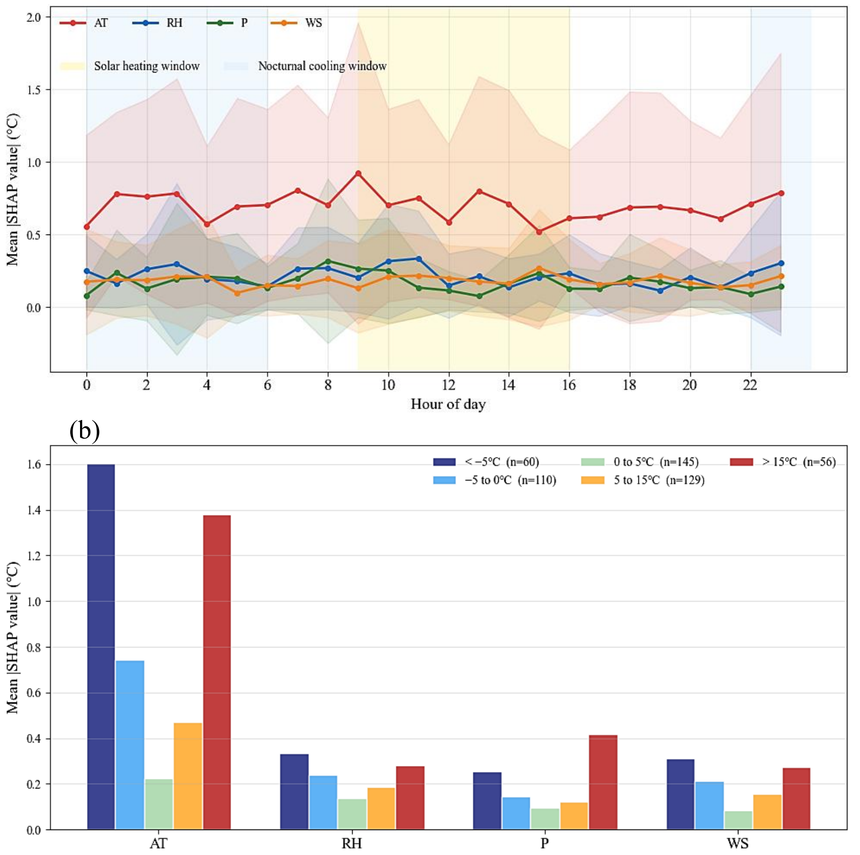


Figure 14. Diurnal variation of mean absolute SHAP values across the 24-hour cycle (a). Shaded regions indicate the solar heating window (09:00-16:00) and nocturnal cooling window (22:00-06:00). Mean absolute SHAP values stratified by RST regime (b). Numbers in parentheses indicate sample counts per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

4.4 Multi-site validation

To assess the cross-site applicability of the proposed ILES framework, validation was conducted at two independent stations: M9474 (Xiaohuangshan, 32.04°N, 119.86°E) and M9448 (Huai'an Airport, 33.75°N, 119.17°E), both located on the central Jiangsu plain. M9474 is situated on a cross-Yangtze River bridge, approximately 363 km southeast of the primary station; M9448 is located near Huai'an Airport, approximately 206 km southeast of M9393, on flat open terrain. As with the primary station M9393, both M9474 and M9448 are situated in relatively unobstructed settings without significant roadside screening structures. At each site, the ILES framework was retrained independently on site-specific observations following the same experimental protocol as described in Sect. 3.3.2. For M9474, the winter of 2024 served as the test set. For M9448, three winter seasons (January–February 2017, December 2017–February 2018, and December 2018–February 2019) were used for training, with December 2019–February 2020 as the test set. The input variables at both stations comprise site-specific air temperature, wind speed, relative humidity, precipitation, and historical RST.~~To assess the cross-site applicability of the proposed ILES framework, validation was conducted at two independent stations: M9474 (Xiaohuangshan, 32.04°N, 119.86°E) and M9448 (Huai'an Airport, 33.75°N, 119.17°E), both located on the central Jiangsu plain. M9474 is situated on a cross-Yangtze River bridge, approximately 363 km southeast of the primary station. M9448 is located near Huai'an Airport, approximately 206 km southeast of M9393, on flat open terrain characteristic of the region; the relatively unobstructed exposure of this site renders it susceptible to stronger advective forcing and greater diurnal RST variability compared to the bridge-mounted station. At each site, the ILES framework was retrained independently on site-specific observations following the same experimental protocol as described in Sect. 3.3.2. For M9474, the winter of 2024 served as the test set. For M9448, three winter seasons (January–February 2017, December 2017–February 2018, and December 2018–February 2019) were used for training, with December 2019–February 2020 as the test set. The input variables at both stations comprise site-specific air temperature, wind speed, relative humidity, precipitation, and historical RST.~~

Table 7 presents the 1-hour prediction results at all three stations. ILES consistently outperforms the standard LSTM model at all three sites, with MAE reductions of ~~17.7~~17.66%, 5.68~~5.7~~%, and 35.65~~35.6~~% at M9393, M9474, and M9448, respectively. The consistent performance advantage of ILES across stations with distinct surface conditions confirms the structural robustness of the proposed approach within the temperate monsoon climate zone.~~The consistent superiority of the ensemble framework across stations with distinct geographical settings and surface conditions confirms the structural robustness of the proposed approach and supports its broader applicability within the study region.~~

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

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

The BRR meta-learner yields closed-form probabilistic forecasts through its posterior predictive distribution (Eq. 20). Fig. 15 presents reliability diagrams evaluated on the held-out test sets at all three stations, in which the empirical coverage curves lie consistently above the diagonal of perfect calibration across all nominal confidence levels, indicating systematic conservative over-coverage at each site. 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, respectively. 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 relative to M9393 reflect the lower RST variability characteristic of the bridge-mounted station, while the consistent coverage hierarchy across all three sites confirms that the conservative calibration pattern is a structural property of the BRR posterior predictive decomposition (Eq. 20) rather than a site-specific artefact.

~~This conservative behavior is operationally preferable for safety-critical winter road maintenance: the cost of under-coverage substantially exceeds that of over-coverage when missing a genuine icing event is considered. In operational deployment, when the lower bound of the 95% predictive interval falls below 0°C, maintenance crews may initiate precautionary pre-treatment irrespective of the point forecast, thereby incorporating quantified uncertainty directly into decision-making. The cross-site consistency of both coverage rates and calibration shape further supports the operational transferability of the probabilistic forecasting component of ILES.~~

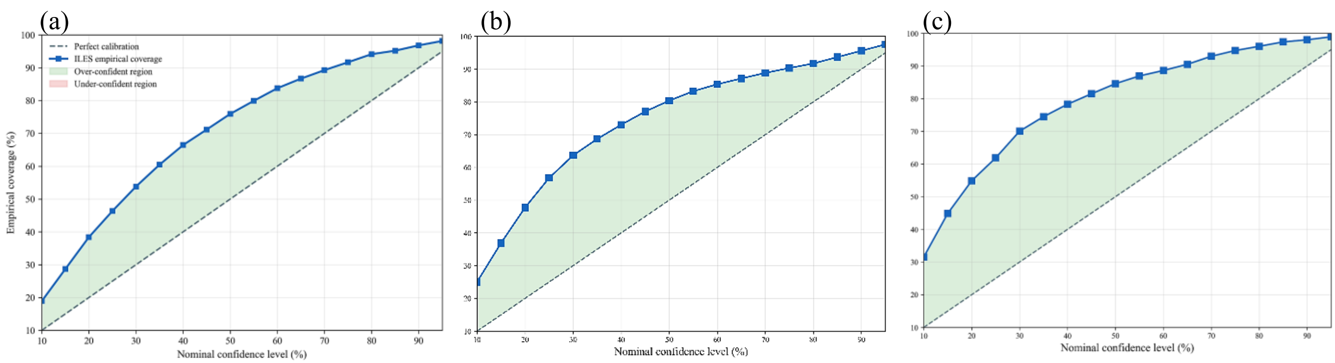


Figure 15. Reliability diagrams of the BRR meta-learner for 1-hour winter RST prediction at M9393, M9474, and M9448.

5 Conclusion

This study proposed the Improved LSTMs Ensemble with Stacking (ILES) framework for winter road surface temperature prediction, integrating two base learners with partially complementary predictive characteristics~~two architecturally complementary base learners~~ within a ~~leakage-free~~ stacking ensemble trained via three-fold temporal cross-validation aligned to complete winter seasons. The framework was evaluated across four analytical perspectives using four consecutive winters of observations from an operational road meteorological station in Jiangsu, China.

On predictive performance, ILES generally achieved the lowest prediction errors among all eleven evaluated models at each of the 1-, 3-, and 6-hour forecasting horizons, with MAE ranging from 0.373°C at 1-hour to 2.108°C at 6-hour and R^2 reaching 0.993, 0.923, and 0.826, respectively. Against the six established machine learning and deep learning baselines—RF, XGBoost, GRU, LSTM, BiLSTM, and CNN-LSTM—ILES reduced MAE by 8.13%–32.67% and RMSE by 8.11%–40.18% at 1-hour, and cut MAE by 5.72%–12.73% and RMSE by 5.41%–16.51% at 3-hour; at 6 hours, ILES achieved lower MAE and RMSE than five of the six established baselines, the exception being GRU, though ILES retained the lowest RMSE across all eleven models. sMAPE rankings were not uniformly optimal: at the 6-hour horizon, GRU, XGBoost, and RF achieved lower sMAPE than ILES, a pattern attributable to the disproportionate sensitivity of sMAPE to near-zero RST values at extended horizons.~~On predictive performance, ILES achieved the lowest errors across all forecasting horizons and evaluation metrics, with MAE ranging from 0.373°C at 1-hour to 2.108°C at 6-hour, and consistent superiority over ten benchmark models spanning persistence, traditional machine learning, and standard deep learning architectures.~~

On input configuration, physics-motivated feature engineering incorporating the surface–air temperature difference and multi-scale RST temporal tendency features overall consistently outperformed both the station-only observational baseline and ERA5-Land reanalysis augmentation across all forecasting horizons and meteorological regimes, including sub-zero, overcast, and clear-sky conditions. These findings demonstrate that meaningful predictive gains at instrumented sites can be achieved through domain-knowledge-guided feature construction alone, without additional sensor infrastructure or external reanalysis data products, and establish Variable combination 3 as the recommended input strategy for operational deployment at monitored road segments.~~On input configuration, physics-motivated feature engineering incorporating the surface-air temperature difference and multi-scale RST temporal tendency features consistently outperformed both the station-only baseline and ERA5-Land reanalysis augmentation across all forecasting horizons and meteorological regimes. These findings demonstrate that meaningful performance gains at instrumented sites can be achieved through domain-knowledge-guided feature construction alone, without additional sensor infrastructure or external data products.~~

On model interpretability, stratified SHAP analysis confirmed that the learned feature importance rankings are qualitatively consistent with the dominant meteorological drivers identified by surface energy balance theory across temperature regimes and the diurnal cycle, with air temperature as the primary predictor and regime-dependent rank shifts among secondary predictors.

On multi-site applicability, ILES maintained its predictive superiority over the LSTM baseline at both independent validation stations within the same temperate monsoon climate zone of Jiangsu Province, with MAE reductions of 17.66%, 5.68%, and 35.65% at M9393, M9474, and M9448, respectively. The BRR meta-learner additionally yielded conservative probabilistic forecasts with empirical prediction interval coverages consistently exceeding nominal confidence levels across all three sites, a calibration pattern confirmed to be a structural property of the BRR posterior predictive decomposition rather than a site-specific artefact.~~On multi-site applicability, ILES maintained its predictive superiority at both independent validation stations with consistent model performance rankings, and the cross-site consistency of BRR calibration further supports the operational transferability of the full probabilistic forecasting framework. The BRR meta-learner additionally yielded conservative probabilistic forecasts with empirical coverages consistently exceeding nominal confidence levels across all three validation sites, a calibration pattern confirmed to be structurally consistent rather than site-specific, and operationally preferable for safety-critical road maintenance decisions.~~

Several limitations motivate future work. The framework was evaluated under a temperate monsoon climate, and its transferability to subarctic, alpine, or maritime regimes remains to be assessed. The absence of direct shortwave radiation measurements and local exposure descriptors—such as sky-view factor, roadside screening geometry, and road orientation—means that the model is best characterized as a station-specific time-series predictor at a fixed exposure.~~The framework was evaluated under a temperate monsoon climate, and its transferability to subarctic, alpine, or maritime regimes remains to be assessed. The absence of direct radiation measurements constrains performance under clear-sky daytime conditions.~~ Integration of numerical weather prediction outputs to extend forecast horizons beyond 6 hours, and the application of physics-informed architectural constraints for data-sparse settings, represent productive directions for future development.

Code and data availability. The codes for conducting the analyses can be downloaded from https://doi.org/10.5281/zenodo.~~19626147~~ 20794320 (Li, 2025). All data used in this study are publicly available.

Author contribution. WTL, LYZ and XHW designed the study. WTL developed the model and wrote the paper. XHW, MMR, YHG, KC and HWL ~~analysed~~analyzed the data. All authors contribute to writing the paper.

Declaration of competing interest. The authors declare no competing financial interests.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (U24A20606, 42575210, 41975087), the Open Research Program of the State Key Laboratory of Severe Weather (2023LASW-B25), and Yunnan Provincial Department of Transportation, Science & Technology – Educational Memorandum No. 152 [2023].

References

Abo-Hashema, M. A.: Modeling pavement temperature prediction using artificial neural networks, in: Airfield and highway pavement 2013: Sustainable and efficient pavements, 490-505, https://doi.org/10.1061/9780784413005.039, 2013.

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

Asefzadeh, A., Hashemian, L., and Bayat, A.: Development of statistical temperature prediction models for a test road in Edmonton, Alberta, Canada, Int. J. Pavement Res. Technol., 10(5), 369-382, https://doi.org/10.1016/j.ijprt.2017.05.003, 2017.

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

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

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

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

Bishop, C. M., Nasrabadi, N. M.: Pattern recognition and machine learning, New York: Springer, https://doi.org/10.1080/15228053.2019.1632410, 2006.

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.

Cheng, H., Liu, J., Sun, L., et al.: Critical position of fatigue damage within asphalt pavement considering temperature and strain distribution, Int. J. Pavement Eng., 22(14), 1773-1784, https://doi.org/10.1080/10298436.2020.1724288, 2021.

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

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

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

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

Darghiasi, P., Zamanian, M., 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.

Gedafa, D. S., Hossain, M., and Romanoschi, S. A.: Perpetual pavement temperature prediction model, Road Mater. Pavement Des., 15(1), 55-65, https://doi.org/10.1080/14680629.2013.852616, 2014.

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.

Ghalandari, T., Shi, L., Sadeghi-Khanegah, F., et al.: Utilizing artificial neural networks to predict the asphalt pavement profile temperature in western Europe, Case Stud. Constr. Mater., 18, e02130, https://doi.org/10.1016/j.cscm.2023.e02130, 2023.

Glorot, X., Bordes, A., and Bengio, Y.: Deep sparse rectifier neural networks, in Proceedings of the 14th Int. Conf. on Artif. Intell. and Statist., JMLR Workshop and Conf. Proc., 315-323, https://doi.org/10.5555/2020408.2020439, 2011.

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

Hassan, H. F., Al-Nuaimi, A. S., Taha, R., et al.: Development of asphalt pavement temperature models for Oman, J. Eng. Res., 2(1), 32-42, https://doi.org/10.24200/tjer.vol2iss1pp32-42, 2005.

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.

Jing, C. and Zhang, J.: Prediction model for asphalt pavement temperature in high - temperature season in Beijing, Adv. Civ. Eng., 2018, 1837952, https://doi.org/10.1155/2018/1837952, 2018.

Joo, C., Park, H., Lim, J., Cho, H., and Kim, J.: Learning-based heat deflection temperature prediction and effect analysis in polypropylene composites using catboost and shapley additive explanations, Eng. Appl. Artif. Intell., 126, 106873, https://doi.org/10.1016/j.engappai.2023.106873, 2023.

Kangas, M., Heikinheimo, M., and Hippi, M.: RoadSurf: a behaviour system for predicting road weather and road surface conditions, Meteorol. Appl., 22(3), 544–553, https://doi.org/10.1002/met.1486, 2015.

Karsisto, V., Nurmi, P., Kangas, M., Hippi, M., and Uppala, A.: Improving road weather model forecasts by adjusting the radiation input, Meteorol. Appl., 23(3), 503–513, https://doi.org/10.1002/met.1574, 2016.

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., Ba, J.: 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.

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

Li, W.: A Hybrid Method for Winter Road Surface Temperature Prediction Using Improved LSTMs and Stacking-Based Ensemble Learning, Zenodo, https://doi.org/10.5281/zenodo.~~19626147~~20794320, 2025.

Li, Y., Chen, J., Dan, H., et al.: Probability prediction of pavement surface low temperature in winter based on bayesian structural time series and neural network, Cold Reg. Sci. Technol., 194, 103434, https://doi.org/10.1016/j.coldregions.2021.103434, 2022.

Li, Y., Liu, L., and Sun, L.: Temperature predictions for asphalt pavement with thick asphalt layer, Constr. Build. Mater., 160, 802-809, https://doi.org/10.1016/j.conbuildmat.2017.11.077, 2018.

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

Liu, B., Yan, S., You, H., et al.: Road surface temperature prediction based on gradient extreme learning machine boosting, Comput. Ind., 99, 294-302, https://doi.org/10.1016/j.compind.2018.03.026, 2018.

Liu, B., Shen, L., You, H., et al.: Comparison of algorithms for road surface temperature prediction, Int. J. Crowd Sci., 2(3), 212-224, https://doi.org/10.1108/IJCS-07-2018-0016, 2018.

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.

Maddu, R., Vanga, A. R., Sajja, J. K., et al.: Prediction of land surface temperature of major coastal cities of India using bidirectional LSTM neural networks, J. Water Clim. Change, 12(8), 3801-3819, https://doi.org/10.2166/wcc.2021.198, 2021.

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

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, 2021b.

Minhoto, M. J. C., Pais, J. C., Pereira, P. A. A., et al.: Predicting asphalt pavement temperature with a three-dimensional finite element method, Transp. Res. Rec., 1919(1), 96-110, https://doi.org/10.1177/0361198105191900112, 2005.

Molavi Nojumi, M., Huang, Y., Hashemian, L., et al.: Application of machine learning for temperature prediction in a test road in Alberta, Int. J. Pavement Res. Technol., 15(2), 303-319, https://doi.org/10.1007/s42947-021-00023-3, 2022.

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

Nantasai, B. and Nassiri, S.: Winter temperature prediction for near-surface depth of pervious concrete pavement, Int. J. Pavement Eng., 20(7), 820-829, https://doi.org/10.1080/10298436.2017.1332466, 2019.

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.

Qiu, X., Hong, H., Xu, W., et al.: Surface temperature prediction of asphalt pavement based on APRIORI-GBDT, in: International Conference on Transportation and Development 2020, 200-212, https://doi.org/10.1061/9780784483183.020, 2020.

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.

Rigabadi, A., Rezaei Zadeh Herozi, M., and Rezagholilou, A.: An attempt for development of pavements temperature prediction models based on remote sensing data and artificial neural network, Int. J. Pavement Eng., 23(9), 2912-2921, https://doi.org/10.1080/10298436.2020.1864011, 2022.

Saliko, D., Ahmed, A., and Erlingsson, S.: Development and validation of a pavement temperature profile prediction model in a mechanistic-empirical design framework, Transp. Geotech., 40, 100976, https://doi.org/10.1016/j.trgeo.2023.100976, 2023.

Schindler, A. K., Ruiz, J. M., Rasmussen, R. O., et al.: Concrete pavement temperature prediction and case studies with the FHWA HIPERPAV models, Cem. Concr. Compos., 26(5), 463-471, https://doi.org/10.1016/S0958-9465(03)00069-7, 2004.

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.

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

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

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.

Tao, R., Peng, R., Wang, H., et al.: Temperature and humidity prediction of mountain highway tunnel entrance road surface based on improved Bi-LSTM neural network, Evol. Syst., 15(3), 691-702, https://doi.org/10.1007/s12530-023-09538-5, 2024.

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.

Wang, D.: Simplified analytical approach to predicting asphalt pavement temperature, J. Mater. Civ. Eng., 27(12), 04015043, https://doi.org/10.1061/(ASCE)MT.1943-5533.0001301, 2015.

Wang, D., Roesler, J. R., and Guo, D. Z.: Analytical approach to predicting temperature fields in multilayered pavement systems, J. Eng. Mech., 135(4), 334-344, https://doi.org/10.1061/(ASCE)0733-9399(2009)135:4(334), 2009.

Wang, X., Pan, P., and Li, J.: Real-time measurement on dynamic temperature variation of asphalt pavement using machine learning, Measurement, 207, 112413, https://doi.org/10.1016/j.measurement.2022.112413, 2023.

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

Wu, Y., Li, C., and Zhang, H.: A review of time-series prediction methods for pavement temperature forecasting, Constr. Build. Mater., 267, 121089, https://doi.org/10.1016/j.conbuildmat.2020.121089, 2020.

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.

Yuan, J., Cheng, H., Sun, L., et al.: Cross-Regional Pavement Temperature Prediction Using Transfer Learning and Random Forest, Appl. Sci., 15(13), 7436, https://doi.org/10.3390/app15137436, 2025.

Zhang, L., Wang, Y., and Chen, J.: Optimization of sliding window parameters for LSTM-based traffic prediction models, IEEE Trans. Intell. Transp. Syst., 23(8), 10698-10708, https://doi.org/10.1109/TITS.2021.3123456, 2022.

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

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

Zhao, X., Shen, A., and Ma, B.: Temperature response of asphalt pavement to low temperatures and large temperature differences, Int. J. Pavement Eng., 21(1), 49-62, https://doi.org/10.1080/10298436.2018.1435877, 2020.