Supplement of Geosci. Model Dev., 19, 9035–9061, 2026
https://doi.org/10.5194/gmd-19-9035-2026-supplement

Geoscientific Model Development
Open Access
EGU

Supplement of

A hybrid method for winter road surface temperature prediction using improved LSTMs and stacking-based ensemble learning

Wanting Li et al.

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

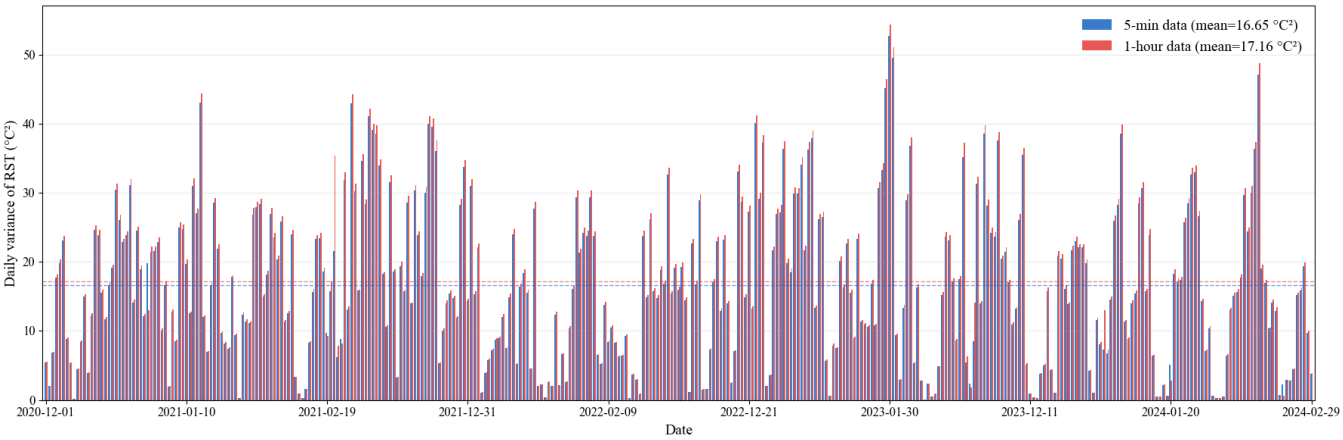

Figure S1. Bar chart of daily variance for the 5-minute raw data and 1-hour data at station M9393. The daily variance of RST was computed separately from the native 5-minute observations (blue) and from the hourly-resampled series (red), for each calendar day across the four winter seasons (December 2020 – February 2024) with complete, unimputed coverage at both resolutions (n = 361 days). Dashed horizontal lines mark the mean daily variance for each resolution (5-minute: 16.65 $°C^2$; 1-hour: 17.16 $°C^2$). The two series track each other closely throughout the study period, with no systematic divergence in any winter season, indicating that hourly aggregation neither suppresses nor inflates genuine day-to-day RST variability.

Table S1. Mean daily variance of RST computed from the native 5-minute series and from the hourly-resampled series, together with their ratio, at all three stations. Here N is the number of calendar days with complete, unimputed coverage at both resolutions. The hourly-to-5-minute variance ratio falls within a narrow band close to unity across all three stations, indicating that hourly resampling introduces only a small, station-invariant smoothing effect and does not materially alter the magnitude of genuine RST variability. This shows that the resampling artefact introduced by aggregating from 5-minute to hourly resolution is not severe.

Station	5-min mean daily variance ($°C^2$)	1-hour mean daily variance ($°C^2$)	N(days)
M9393	16.650	17.164	361
M9474	6.856	7.042	361
M9448	19.380	19.925	330

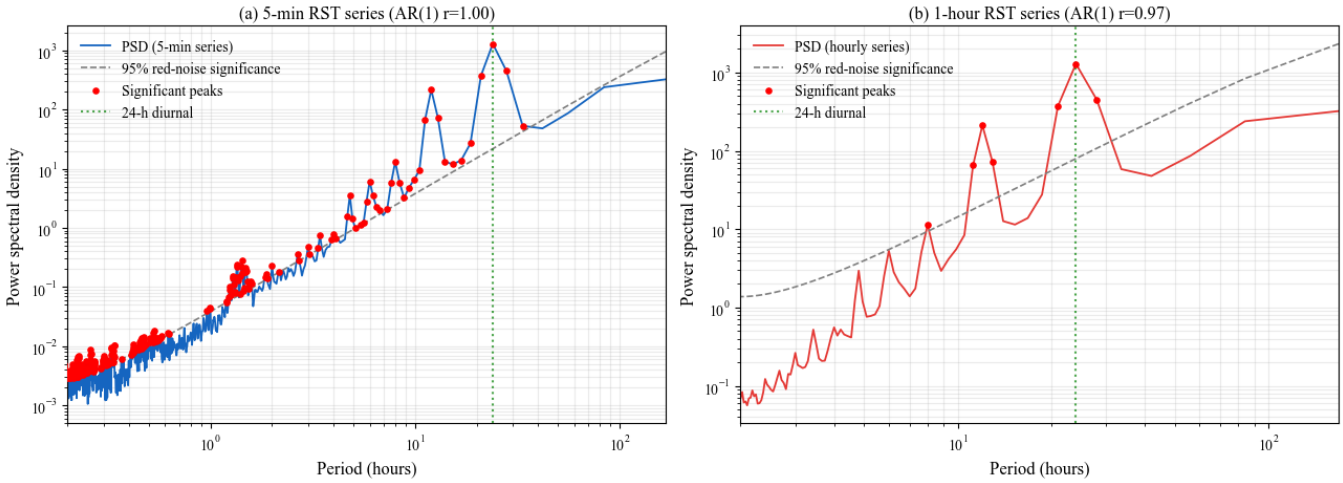

Figure S2. Power spectral density (PSD) of RST at station M9393, computed from (a) the native 5-minute series and (b) the hourly-resampled series, over the longest continuous winter segment common to both resolutions (2,184 hours). Statistical significance of spectral peaks was assessed against a first-order autoregressive (AR(1), "red noise") background spectrum at the 95% confidence level (dashed line); filled red circles mark significant peaks. The AR(1) coefficient is similarly high at both resolutions (5-minute: r = 1.00; 1-hour: r = 0.97), reflecting the strong autocorrelation of RST at both timescales. The dominant significant periodicities — the 24-hour diurnal cycle (dotted vertical line) and its 12-hour semi-diurnal harmonic — are identical between the two resolutions, and no spurious low-frequency period appears in the hourly series that is not already present in the 5-minute series. This is the diagnostic signature expected in the absence of

aliasing: hourly block-averaging acts as an inherent low-pass (anti-aliasing) filter that preserves the genuine periodic structure of RST rather than folding unresolved high-frequency content into spurious low-frequency peaks.

Table S2. MAE, RMSE, and sMAPE improvement of ILES relative to the established machine learning and deep learning baselines (RF, XGBoost, GRU, LSTM, BiLSTM, CNN-LSTM), across forecasting horizons. For each horizon and metric, "Baseline range (best–worst)" reports the best- and worst-performing established baseline (named in parentheses) and its metric value; "ILES" gives ILES's own value; "Improvement range" is the percentage error reduction achieved by ILES relative to that baseline range, computed as (baseline − ILES) / baseline × 100%. At the 1- and 3-hour horizons, ILES achieves a lower error than every established baseline on every metric (improvement range entirely positive). At the 6-hour horizon, ILES's MAE and sMAPE fall within rather than below the baseline range (negative lower bound: −5.03% for MAE, −5.79% for sMAPE, both attributable to GRU), while its RMSE remains the lowest among all evaluated models. This table is the quantitative basis for the horizon- and metric-qualified superiority claims made for ILES in Sect. 4.1.

Horizon	Metric	Baseline range (best–worst)	ILES	Improvement range
1-hour	MAE	0.406 (CNN-LSTM)–0.554 (XGBoost)	0.373	8.13%–32.67%
	RMSE	0.567 (CNN-LSTM)–0.871 (XGBoost)	0.521	8.11%–40.18%
	sMAPE	21.41% (CNN-LSTM)–26.07% (XGBoost)	20.33%	5.14%–22.02%
3-hour	MAE	1.345 (GRU)–1.453 (LSTM)	1.268	5.72%–12.73%
	RMSE	1.940 (GRU)–2.198 (LSTM)	1.835	5.41%–16.51%
	sMAPE	50.97% (XGBoost)–54.89% (LSTM)	47.55%	6.71%–13.38%
6-hour	MAE	2.007 (GRU)–2.366 (LSTM)	2.108	-5.03%–10.90%
	RMSE	2.792 (GRU)–3.452 (LSTM)	2.754	1.36%–20.22%
	sMAPE	67.38% (GRU)–75.00% (CNN-LSTM)	71.28%	−5.79%–4.95%

Table S3. Paired t-tests comparing the absolute forecast error of ILES against each of its two constituent base learners, computed across all test-set samples at the 1-hour forecasting horizon. Mean difference in this table refers to the sample-average difference in absolute forecast error, ILES minus baseline, across all test samples; a negative value indicates ILES has lower mean absolute error. Both comparisons are statistically significant at $p < 0.001$, with small-to-moderate effect sizes (Cohen's $d = -0.105$ for ILES vs KNN-LSTM; −0.166 for ILES vs BiLSTM-MHA), consistent with a real but modest error reduction from the stacking step rather than a null or a large effect.

Comparison	Mean difference (℃)	t-statistic	p-value	Cohen's d
ILES vs KNN-LSTM	−0.0163	$t = -4.899$	< 0.001	−0.105
ILES vs BiLSTM-MHA	−0.0198	$t = -7.702$	< 0.001	−0.166