Response to Reviewer 1

Dear Reviewer:
We sincerely thank the Reviewer for the positive assessment and for confirming that the numerical entries in Tables 4–7 are internally consistent. We are grateful for the recommendation of acceptance and address the three proof-stage suggestions below.

1. *"Cross-references to the Supplement. The evidence for (i) the 1-min predictive skill being robust to resampling artefacts (Figs. S1–S2) and (ii) the pairwise significance tests supporting the rankings in Tables 5–7 (Tables S3–S6) is convincing. For reader convenience, please consider adding explicit pointers from the main text — a parenthetical (Supplement Figs. S1–S2) next to the 1-min robustness statement in Sect. 3.2, and a short note in the captions of Tables 5–7 referring to Supplement Tables S3–S6."*

Response:
We agree and have added these pointers, with one clarification on scope. Checking the Supplement material against what each table actually contains: Supplement Figure S1 (Bar chart of daily variance for M9393 site 5-minute raw data and 1-hour data), Figure S2 (Mean daily variance of RST computed from the native 5-minute series and the hourly-resampled series, together with their ratio, at all three stations)—referred to as Figs. S1–S2 in the previous response letter—and Table S1 (mean daily variance ratio and matching periodogram peaks between the 5-minute and hourly series)—referred to as Table S3 in the previous response letter—together support the resampling-robustness statement now in Sect. 3.1.1. We have added the parenthetical citation there, immediately after the sentence describing hourly aggregation: "…The resulting hourly series exhibits a near-unity variance ratio relative to the native 5-minute series and matching significant periodicities (Supplement Figs. S1–S2 and Table S1)."

Supplement Table S2 (MAE, RMSE, and sMAPE improvement of ILES relative to the established machine learning and deep learning baselines across forecasting horizons) — referred to as Table S5 in the previous response letter — and Table S3 (Paired t-tests comparing the absolute forecast error of ILES against each of its two constituent base learners) — referred to as Table S6 in the previous response letter — support specifically the model rankings reported in Table 5 (Sect. 4.1); we have added the note "Pairwise significance testing supporting the performance rankings reported in this table is provided in Supplement Tables S2 and S3" to Table 5's caption.

We have not added an equivalent note to the captions of Table 6 (input-variable-combination comparison) or Table 7 (multi-site validation), because no pairwise significance testing was conducted for either of those two comparisons — Tables S2 and S3 are both specific to the eleven-model comparison in Table 5 and do not extend to the three-configuration comparison in Table 6 or the five-model, three-station comparison in Table 7.

2. *"Abstract, L32. outperform → outperforms."*

Response:
We appreciate the Reviewer for pointing out this grammatical error, and the sentence has been revised accordingly. "physics-motivated feature engineering … outperform both the station-only baseline…" now reads "…outperforms both the station-only baseline…", agreeing with the singular subject.

3. *"Spelling consistency. The manuscript mixes British and American spellings in a few places (e.g. modelling/modeling, characterise/*characterize). A light pass to unify one convention, per the journal's house style, would be helpful."*

Response:
We thank the Reviewer for this observation. We have performed a full-text pass and unified all spelling to British English, which was already the predominant convention used throughout the manuscript (e.g., "modelling", "characterise", "normalisation", "optimise", "analyse").

Manuscript revisions

The following American-spelling instances have been corrected to their British equivalents throughout the manuscript and captions: “modeling” → “modelling” (5 instances), “characterize” → “characterise” (3 instances), “optimization” → “optimisation” (2 instances), “normalization” → “normalisation” (1 instance), “behavior” → “behaviour” (2 instances), “analyze” → “analyse” (1 instance). No further mixed-spelling instances remain following this pass.

Response to Reviewer 2

Dear Reviewer:

We thank the Reviewer again for the exceptionally careful, forensic reading of the preprocessing pipeline and of the consistency between our previous response letter and the revised manuscript. We agree that several items were not fully reconciled in the previous round, and we address each of the nine points below individually, in the order raised.

1. *"The manuscript at line 343 states that gaps not exceeding two consecutive hours were filled by linear interpolation, whereas the leakage audit in the response letter assumes a threshold of one consecutive hour throughout. Please state the threshold that was implemented. If it is two hours, please repeat the cross hour anchor audit at that threshold, since a gap of that length can span a full clock hour and the interpolation anchor is then necessarily drawn from a later hour."*

Response:

We thank the Reviewer for catching this. The threshold actually implemented and audited — in the previous round and in every reported result — has been one consecutive hour (⩽12 successive missing 5-minute slots), not two. Line 343 belonged to a paragraph of Sect. 3.1.1 that had been deleted (struck through in red) during the previous revision and, in the manuscript version the Reviewer examined, had not yet been replaced by the corrected paragraph given under Comment 5; this made the manuscript appear to state a 2-hour threshold when the actual computation, and our previous response letter, both used 1 hour throughout.

Manuscript revision

The consolidated Sect. 3.1.1 paragraph (reproduced in full under Comment 5 below) now states "gaps not exceeding one consecutive hour (≤ 12 successive missing 5-minute slots)" consistently.

2. *"The long gap fill is described in the manuscript as a training period climatological hourly mean only, while the response letter describes it as that mean combined with spatiotemporal information from neighbouring stations. Please reconcile these. If neighbouring station data were used, please specify the procedure and the time stamps involved, since these are not a fixed constant and may introduce a look ahead path."*

Response:

We apologise for this inconsistency, and we would like to clarify how it arose. In responding to the previous round's concern about the leakage-robustness of our imputation procedure, we surveyed the literature on long-gap imputation methods for meteorological and road-weather time series, and found that published approaches broadly fell into two families: (i) a station-specific climatological mean, using only that station's own historical record at matching calendar times [cite: e.g., Yin et al., 2019; Darghiasi et al., 2023], and (ii) methods that additionally draw on spatiotemporal information from neighbouring stations, typically within a spatial interpolation or network-based road-weather model [cite: e.g., Crevier and Delage, 2001; Karsisto, 2024]. We cited representative examples of both in the response letter accompanying that round, intending this as a literature survey establishing that our chosen approach is a recognised, standard method — not as a description of a hybrid procedure implemented in this study.

On rereading that response letter, we recognise that the surrounding text did not clearly separate "methods described in the literature" from "the method actually implemented here", and the neighbouring-station approach was consequently read as part of our own procedure. This was an error in how the literature review was presented, not a discrepancy between what was computed and what was reported: the manuscript's description — a station-specific climatological mean, computed solely from that station's own training-period observations — is, and has always been, the only long-gap imputation method implemented in this study. We have revised the response letter to state this distinction explicitly, and we confirm that no neighbouring-station data of any kind were used at any stage of the preprocessing pipeline for any of the three stations.

The long-gap fill value for each missing 5-minute slot is computed as the mean of the observed values recorded at the same hour-of-day and same calendar period across the three training winters (December 2020 – February 2021, December 2021 – February 2022, and December 2022 – February 2023), and is applied identically regardless of whether the corresponding gap occurs in the training or test period. Because this value is derived exclusively from training-period observations at matching

calendar times from three prior winters, it introduces no look-ahead information from any test-period observation, irrespective of the gap's position in the series.

Manuscript revision

Sect. 3.1.1 has been revised to state this procedure precisely (see the consolidated paragraph under Comment 5).

References in support of this response:

Crevier, L. P., 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.

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/jcrgei.creng-691, 2023.

Karsisto, V. E.: RoadSurf 1.1: open-source road weather model library, Geosci. Model Dev., 17(12), 4837-4853, https://doi.org/10.5194/gmd-2023-175, 2024.

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.1729, 2019.

3. *"The stated order of preprocessing operations places imputation before the chronological train and test split, while the accompanying text asserts that interpolation uses only within period observations. These cannot both hold. Please state the order that was implemented and make the manuscript consistent with it."*

Response:

We agree this was a genuine inconsistency in the previous revision's description (Steps 1–2 of our prior response letter performed imputation before Step 3's train/test split, while the same response letter asserted within-period-only interpolation). To remove this ambiguity, we have re-implemented the pipeline so that the chronological train/test split is performed immediately after outlier detection and before any imputation step, for all three stations. The corrected order is: (1) chronological train/test split (training: December 2020–February 2023; test: December 2023–February 2024); (2) physical-range outlier detection on the full 5-minute series; (3) within each period independently, short-gap linear interpolation using only that period's own observations, and long-gap climatological filling using the training-period-only constant described in our response to Comment 2 above; (4) within each period independently, hourly aggregation; (5) fold construction for the out-of-fold stacking procedure (Sect. 2.3), applied to the training-period hourly data only.

Manuscript revision

Sect. 3.1.1 has been rewritten to state the corrected order explicitly (see the consolidated paragraph under Comment 5), replacing the previous ambiguous description. The archived dataset at https://doi.org/10.5281/zenodo.22020890 has been updated to include, for each station, the independently processed {STATION}_training_set_cleaned.csv and {STATION}_test_set_cleaned.csv files, so that the split-first processing order is independently verifiable from the archived data itself.

4. *"Please restore to Section 3.1.1 the aggregation detail deleted at lines 346 to 348, namely that precipitation was aggregated as hourly totals and all other variables as hourly means, together with the rule that hours with fewer than six valid five-minute records were treated as missing. The latter does not currently appear anywhere in the manuscript, and without both a reader cannot reconstruct the hourly dataset."*

Response:

We apologise for this omission: in condensing Sect. 3.1.1 during the previous revision, we deleted the aggregation-rule sentence rather than only the superseded threshold/order language around it. Both details have been restored to Sect. 3.1.1.

Manuscript revisions

Section 3.1.1 has been rewritten in full to incorporate the resolutions to points 1–4 above (see the consolidated paragraph under Comment 5).

5. *"Please state in the manuscript, rather than only in the response letter, the fraction of test period hourly values that incorporate an interpolation anchor from a later clock hour, and document how imputed values were excluded from the evaluation targets."*

Response:

We agree and have consolidated Comments 1–5 into a single revised Sect. 3.1.1 paragraph, reproduced below, which states the split-first order, the corrected 1-hour threshold, the cross-hour-anchor mechanism, the reconciled long-gap procedure, the restored aggregation rule, and the resampling-robustness cross-reference (Reviewer 1, Comment 1) in one place. Because M9474 and M9448 are introduced later in the manuscript as multi-site validation stations (Sect. 4.4), we report their audit results there rather than in Sect. 3.1.1, alongside the description of their retraining protocol, so that each station's imputation statistics appear where that station is first discussed.

We have additionally added, in Sect. 3.1.3 (Evaluation index), the sentence: "All reported evaluation metrics were computed exclusively against directly observed RST values; imputed hourly values were retained as model inputs where they occur but were excluded from the evaluation targets."

Revised Sect. 3.1.1 paragraph:

"Data quality control was applied in several steps. First, 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. Physically implausible values were then rejected as sensor errors: RST observations exceeding 50 ℃ or falling below −40°C, negative precipitation values, and relative humidity outside the range [0%, 100%] were removed. For gaps not exceeding one consecutive hour, linear interpolation was applied using the nearest valid observations bracketing the gap, drawn only from that same period. If the right-side interpolation anchor falls within the same natural hour as the missing slot(s), the resulting hourly mean incorporates no observation from any later hour, and no cross-hour information enters the model by construction.

Longer gaps, exceeding one consecutive hour and occurring primarily during scheduled sensor maintenance, were filled using the mean of the observed values recorded at the same hour-of-day and same calendar period across the three training winters, and applied identically regardless of whether the corresponding gap occurred in the training or test period; because this value is derived exclusively from training-period observations at matching calendar times from three prior winters, it introduces no look-ahead information from any test-period observation. This long-gap climatological fill accounts for 1.18% of 5-minute test-period samples at M9393. Following imputation, all variables were resampled to hourly resolution within each period: precipitation was aggregated as the hourly total, and all other variables were computed as hourly means to suppress transient sub-hourly fluctuations. The resulting hourly series exhibits a near-unity variance ratio relative to the native 5-minute series and matching significant periodicities (Supplement Figs. S1–S2 and Table S1). This yielded a final dataset of 8,664 samples for model development."

Supplement to Section 4.4:

"…Consistent with the imputation procedure described in Sect. 3.1.1, no short-gap event in the test period has a cross-hour interpolation anchor at either validation station, and long-gap climatological fills — computed using the same station-specific, training-period-only procedure — account for 0.66% and 0.13% of 5-minute test-period samples at M9474 and M9448, respectively…"

6. *"The phrase "leakage-free stacking ensemble" appears to remain in the Abstract at line 21, although it has been removed from the Introduction and Conclusion and the response letter states it was removed from the Abstract. Please remove it there as well."*

Response:

We thank the Reviewer for this observation, and we would like to clarify what we believe happened here, since on inspection the phrase "leakage-free stacking ensemble" does not appear anywhere in the current Abstract text.

This exact phrase was removed from the Abstract in an earlier revision round, prior to the version most recently submitted. The file the Reviewer examined was a tracked-changes (marked-up) version of the manuscript; it is possible that this specific

deletion — made in an earlier round — had not been "accepted" in Word's track-changes history and therefore still appeared as struck-through text in that file, which could reasonably be read as the phrase still being present. We apologise for this: it reflects an artefact of our revision file hygiene rather than the phrase having survived unaddressed in the actual manuscript content.

To remove any ambiguity, we confirm that the current Abstract sentence at L21 reads in full: "This study proposes the Improved LSTMs Ensemble with Stacking (ILES) framework, which integrates two base learners with partially complementary predictive characteristics within a stacking ensemble employing out-of-fold cross-validation." Neither "leakage-free" nor any other unscoped leakage-related claim appears in this sentence or elsewhere in the Abstract. We have additionally re-exported the tracked-changes file for this round with all changes from every prior round accepted, so that only the current, final wording is visible, and no historical deletion can be mistaken for outstanding text.

7. *"Two uncertainty claims remain unscoped: line 162, which refers to closed form uncertainty quantification as the final ensemble output, and line 290, which describes probabilistic outputs quantifying prediction uncertainty. Please apply the same meta learner scoping used elsewhere."*

Response:

We agree and have revised both instances to match the scoping already applied.

Manuscript revisions

L157 (previously L162): "…yielding the final ensemble output." has been revised to "…yielding probabilistic forecasts with closed-form uncertainty quantification as the final ensemble output." L282 (previously L290): "⋯BRR yields probabilistic outputs that quantify prediction uncertainty⋯" has been revised to "⋯BRR yields probabilistic outputs that quantify prediction uncertainty at the ensemble combination layer, conditional on the deterministic base-learner outputs⋯" We have additionally searched the full manuscript for any other unscoped instance of "uncertainty quantification" and confirm none remain.

8. *"The following also do not appear: the second half of the promised Section 4.4 scope sentence regarding transferability across stations with broadly similar exposure geometries; the third sentence of the promised Conclusion limitations paragraph regarding extrapolation to nearby road segments; the forward reference to the sub additive ablation result promised for the Introduction; and the removal of "architectural complementarity" at line 485."*

Response:

We thank the Reviewer for this careful check. Upon comparing the file submitted for the previous round against our internal revision log, we confirm that all four items were correctly drafted during the previous revision but were inadvertently left out of, or truncated within, the manuscript file that was submitted. We have restored all four in full and independently verified their presence in the current manuscript.

Manuscript revisions

(a) Sect. 4.4, opening paragraph, now reads in full: "As with the primary station M9393, both M9474 and M9448 are situated in relatively unobstructed settings without significant roadside screening structures; this validation therefore assesses transferability of the framework's architecture and training protocol across stations sharing broadly similar exposure geometries, rather than across the specific shading or sky-view conditions encountered at screened road segments."

(b) The Conclusion's limitations paragraph now includes the third sentence: "The learned diurnal pattern in the RST input sequence is conditioned on the specific radiation environment at each monitored cross-section and should not be extrapolated to nearby road segments with substantially different shading or orientation without site-specific retraining or the inclusion of explicit radiation inputs."

(c) The Introduction now includes, following the description of the stacking ensemble: "⋯within a stacking ensemble⋯; the ablation analysis shows that the combined gain from these two components is sub-additive."

(d) L463's (previously L485) "architectural complementarity" has been replaced with "partially complementary predictive characteristics", consistent with the terminology standardised throughout the rest of the manuscript in the previous round.

9. *"Line 716 states that Variable combination 3 shows consistent predictive superiority across all forecasting horizons and meteorological regimes, and line 827 retains equivalent phrasing with the word "overall" inserted. Table 6 shows that its 6 hour sMAPE of 100.355 percent is the highest of the three configurations. Please qualify both in the same manner already applied to Sections 4.1 and 4.2."*

Response:

We agree and thank the Reviewer for catching that this qualification, applied in Sects. 4.2 and 5 during the previous round, was not carried through to L660 (previously L716) and L755 (previously L827).

L660 (previously L716) read: "Variable combination 3 is recommended as the preferred input strategy given its consistent predictive superiority demonstrated across all forecasting horizons and meteorological regimes." This has been revised to: "Variable combination 3 is recommended as the preferred input strategy given its consistently lowest MAE and RMSE across all forecasting horizons and meteorological regimes."

L755 (previously L827) has been revised from "⋯physics-motivated feature engineering⋯overall consistently outperformed both the station-only observational baseline and ERA5-Land reanalysis augmentation across all forecasting horizons and meteorological regimes" to: "⋯physics-motivated feature engineering⋯achieved the lowest MAE and RMSE relative to both the station-only observational baseline and ERA5-Land reanalysis augmentation across all forecasting horizons and meteorological regimes examined."

We have additionally re-searched the full manuscript for any other instance of "consistent⋯superiority" or similar absolute phrasing applied to Variable combination 3 and confirm none remain unqualified.

We believe these revisions fully resolve the nine remaining points and the three proof-stage suggestions, and we are prepared to provide a clean, corrected manuscript together with the revised Supplement for the Reviewers' consideration.