

Atmospheric Data Science Letters (demo journal)

Manuscript ID: ADSL-2026-0917

Title: AttnLite-LSTM: A Lightweight Attention-Enhanced LSTM for 24-Hour PM2.5 Forecasting Across Twelve Chinese Cities

Decision: Major Revision

Date: 15 October 2026

Editor's Comments

Dear Author(s),

Your manuscript has been evaluated by two reviewers. Both find the topic relevant, but Reviewer 1 raises serious concerns about the experimental design that must be addressed before the paper can be considered further. Please provide a point-by-point response to all comments and highlight the changes in the revised manuscript.

E1. In addition, please include a Data and Code Availability statement as required by the journal policy.

E2. The manuscript would benefit from careful language editing.

Reviewer #1

The authors propose an attention-enhanced LSTM for PM2.5 forecasting. While the paper is clearly written, I have major concerns about the validity of the evaluation.

1. Section 2.4 states that samples were *randomly* split into training, validation and test sets. For hourly time series with a 48-h sliding window, this causes severe information leakage, because adjacent overlapping windows end up in both the training and test sets. The reported improvements may therefore be overestimated. The authors must use a chronological split (e.g., train on 2018–2020, validate on 2021, test on 2022) or a rolling-origin evaluation and re-run all experiments.

2. Section 2.2: min–max normalisation is computed on the full dataset, i.e., including the test period. Scaling parameters should be estimated on the training set only.

3. The novelty is limited. Attention-based LSTMs for air-quality forecasting have been published many times. The baselines (ARIMA, XGBoost, LSTM, GRU) are not state of the art. Please compare against at least one existing attention-LSTM model and recent Transformer-based forecasters such as Informer or PatchTST, and clarify what is genuinely new.

4. No statistical test is provided to show that the improvements are significant, and results appear to be from a single run. Please report mean $\pm$ standard deviation over multiple random seeds and apply, e.g., the Diebold–Mariano test.

5. Section 2.2: linear interpolation is reasonable for short gaps but not for gaps of several days. Please report the distribution of gap lengths and explain how long gaps were handled.

6. How were the hyperparameters of the baseline models selected? A fair comparison requires the same tuning budget for all models.

7. Section 3.3 claims that the attention weights “explain which inputs drive the model's

predictions”. It is well known that attention weights are not necessarily faithful explanations. Please soften this claim or support it with additional evidence (e.g., perturbation or ablation analysis).

Reviewer #2

This is a practically useful study with a large multi-city dataset. I have the following suggestions.

1. Public-health warnings matter most during heavy pollution episodes. Please evaluate performance separately for hours with $\text{PM}_{2.5} > 75 \mu\text{g}/\text{m}^3$ and $> 150 \mu\text{g}/\text{m}^3$.
2. Only metrics averaged over 24 forecast hours are reported. Please show how the error changes with lead time (1–24 h).
3. Section 2.3: do the inputs include meteorological observations from the forecast period? In operational settings future meteorology is unknown, so using observed future meteorology would be unrealistic. Please clarify.
4. Section 3.4: the efficiency claim lacks context. Please specify the hardware and report the number of parameters and inference time for all models.
5. Minor issues: (a) Eq. 4 uses H while Eq. 1–3 use h_t ; (b) units are missing in Table 2; (c) Figure 2 has low resolution; (d) the 19.6% improvement in Section 3.2 should state whether it refers to RMSE averaged over cities.
6. The statement that this is “the first lightweight attention model for multi-city $\text{PM}_{2.5}$ forecasting” (Abstract and Introduction) is difficult to verify and should be toned down.
7. The Discussion is very short. Please discuss the poorer performance in Shenzhen and the limitations of the study in more depth.