Review of "A Physics-Constrained Deep-Learning Framework for Retrieving Vertical Distribution of PM2.5 Chemical Components" (AMT)

This manuscript presents a novel lidar-based retrieval framework that integrates deep learning with physics-constrained optimization to estimate vertical mass concentration profiles of five $PM_{2.5}$ chemical components (SO_4^{2-}, NO_3^-, NH_4^+, OM, BC). The topic is scientifically important: retrieving aerosol composition profiles from lidar would significantly advance air quality monitoring, chemical transport modeling, and source apportionment. The combination of physics constraints and deep learning is innovative and promising.

However, while the conceptual idea is strong, the manuscript lacks clarity in describing the model framework and provides insufficient evidence that the approach accurately captures the physics of vertical aerosol composition or generalizes across seasons, sites, and aerosol regimes. Significant issues in methodology, validation, and presentation hinder the scientific interpretation of the results. I therefore recommend **major revision**.

Major Concerns

1. Overall framing, workflow clarity, and Figure 1

- **Figure 1 is difficult to interpret**: inputs/outputs are not clearly labeled, colored boxes lack explanation, and several acronyms are undefined. The figure should be redesigned as a **clear block-flow diagram** that lists:
  - all inputs (with units, vertical resolution, and dimensionality),
  - each module's output,
  - loss functions used,
  - data flow direction and optimization loops.
- **Reorder sections** so the Data section precedes the Model description. Readers must understand what data the model consumes before interpreting architectural choices.

2. Ambiguity in algorithm description

- Section 2.1.1 is **confusing and lacks foundational background**, making the workflow difficult to follow without jumping back and forth.
- It is unclear **what the deep-learning model predicts per vertical level**. Please explicitly specify:
  - whether the model outputs component concentrations, component fractions, categorical flags, or something else,
  - the *exact dimensionality* (e.g., levels × 5 components).
- Clearly define the **target variables** and how they are constructed.
- Provide detailed descriptions of the **multi-objective optimization**, including:
  - inputs and outputs,
  - spatial/temporal/vertical resolution,
  - how physics constraints are incorporated mathematically.
- The purpose of using mentioned components/models.

- The rationale for using a **two-step prediction process** (component "flags" followed by concentrations) rather than a single multi-output network is not explained. The manuscript would benefit from an experimental justification or comparison.

3. Temporal and spatial data splits

- The current **random 80/20 split** is not appropriate for meteorological/aerosol time series due to temporal autocorrelation, which risks information leakage.
- Consider implementing:
    - **temporal holdouts** (e.g., full seasons),
    - **spatially independent test sites**,
    - **blocked k-fold cross-validation** preserving temporal/spatial independence.
- The manuscript evaluates an *independent* dataset only in the Results section, but this dataset should be partially used for the validation/testing framework.
- The reported error statistics for the independent dataset are **not clearly presented and differ considerably** from training results. For a well-generalized model, **validation and independent-test errors should be similar**; their discrepancies raise concerns about generalization and physical consistency.
- Surface-only scatterplots from the training year are insufficient to establish model validity, especially given that the model's primary output is a **vertical distribution**.

4. Weak vertical-profile validation

The manuscript focuses on retrieving vertical composition profiles but presents **minimal validation** of these profiles.

I strongly recommend including:

- Direct comparisons with **aircraft or in situ vertical measurements**, using metrics such as bias, RMSE, MAE, percent error, and correlation *at each altitude bin*.
- **Case studies** across representative aerosol regimes (smoke, dust, pollution, background).
- Aggregated statistics by:
    - altitude,
    - site,
    - aerosol type,
    - season.

If vertical observational data are limited, the manuscript should **explicitly quantify these limitations** while still presenting as much vertical validation as possible.

5. Heterogeneous site performance

- Figure 5b shows substantial site-to-site variability: some sites have nearly zero correlation, while the best site reaches ~0.6.
- Please investigate and report potential causes, such as:
    - aerosol-type mismatch,

    - representativeness of training data,
    - site-specific meteorology or emissions,
    - instrument characteristics.
- Consider:
    - a map showing training vs. test sites,
    - per-site metrics (MAE, RMSE, bias, percent error, N),
    - problematic site scatterplots or boxplots to illustrate error spread.

6. Lack of uncertainty quantification

Given the physics-constrained framing, the model should also provide **uncertainty estimates**, or at minimum a discussion of uncertainty propagation. Possible approaches include: Ensemble modeling, Monte Carlo dropout, error propagation from lidar extinction + physics constraints. Uncertainty bounds would greatly strengthen confidence in profile retrievals.

Minor Comments and Suggestions

1. If possible, include an **ablation study** comparing architectures (CNN, BiLSTM, CNN+BiLSTM, transformer) to justify the chosen hybrid design.
2. Clarify the meaning of "estimated and observed extinction coefficients" (line 192). Does "estimated" refer to IMPROVE-derived extinction?
3. Define **all acronyms** at first use; ensure figure captions are self-contained.
4. The manuscript describes z-score normalization but not how **denormalization** is performed. Why use aircraft-based measurements for denormalization instead of lidar-derived extinction? Explain and quantify the sensitivity.
5. The scaling procedure using the ratio of in situ to aircraft $PM_{2.5}$ ("initially scaled…") is ambiguous. Provide a **clear mathematical expression** and discuss whether this introduces bias.
6. The description of the **attention layer** lacks physical interpretation. Is attention purely data-driven, or does physics guide attention weights? If physics influences attention, show how.
7. Figure 1 needs explicit legends for color boxes/arrows and clear annotation of all inputs and outputs.
8. Figure 5a does not effectively show differences between datasets. Consider:
    - scatterplots colored by site with standard deviations,
    - an additional plot showing error distribution histograms for each of the five components.
9. Include full **training hyperparameters**: batch size, learning rate, optimizer, epochs, early stopping criteria, normalization statistics.