

Response to Reviewers'

Comments

Reviewer 3

Overall evaluation

Reviewer comment (in blue)

Overall evaluation

This manuscript addresses cloud detection in multi-wavelength polarized Raman lidar networks by developing a three-stage framework that combines self-supervised pre-training, VDE-based empirical prior distillation, and limited expert-annotated fine-tuning. The framework is evaluated on the CARLNET dataset under multi-site and cross-temporal conditions. The study addresses important operational challenges, including incomplete near-field overlap, weak upper-level cloud signals, varying signal-to-noise ratios, energy fluctuations, and aerosol–cloud discrimination. The methodology is systematically designed, and the independent cross-site/cross-temporal evaluation, together with radiosonde comparisons, provides a valuable assessment framework.

However, several aspects require further clarification and revision. The novelty of the framework should be described more carefully, as the main contribution appears to be the integration of existing machine-learning and physical-guidance approaches rather than a fundamentally new cloud identification principle. Several methodological choices require stronger justification, the label-generation process needs greater transparency, potential biases introduced through VDE-based distillation should be discussed, and the statistical robustness of the reported improvements requires further evaluation. Some conclusions also appear stronger than supported by the current evidence. I recommend major revisions addressing these methodological and interpretational issues before publication.

Author response (in black)

We thank the reviewer for the systematic assessment. We made focused revisions regarding the framing of novelty, calibration dependence, dataset splitting,

methodological parameters, label reliability, potential VDE bias, the evidential role of radiosonde comparisons, and uncertainty in the reported results. We also moderated conclusions that extended beyond the evidence provided by the three independent CARLNET test sites.

Location of revision

Introduction, Lines 94–118; Sections 2.1–2.6, Lines 120–304; Sections 3.1–3.4, Lines 322–436; Sections 4.1–4.3, Lines 453–656; Conclusion and Discussion, Lines 658–711; Appendix A, Lines 714–860.

Major comment 1 Novelty and calibration wording

Reviewer comment

1. The novelty should be framed more accurately

The manuscript presents the framework as “calibration-decoupled” and emphasizes the combination of self-supervised learning and VDE-guided distillation. While the proposed framework is well motivated and practically valuable, its primary novelty appears to be the integration of existing concepts into a lidar-network-specific workflow rather than the introduction of a fundamentally new cloud detection principle. The term “calibration-decoupled” may also be overstated. Although the method reduces dependence on fully calibrated optical retrieval products, it still relies on processed lidar observables and VDR information, which requires polarization gain-ratio calibration. Therefore, the framework is not entirely independent of calibration. The authors should consider using more precise terminology.

Author response

We agree with the reviewer's assessment and have redefined the scope of the novelty claimed in this study. This study does not propose a new cloud-scattering mechanism or a fundamentally new physical criterion for cloud detection. Its methodological contribution lies in integrating self-supervised structural learning from unlabeled observations, probabilistic weak supervision provided by VDE, and fine-tuning with limited expert annotations into a unified cross-site workflow for operational Raman lidar network observations. The contribution therefore primarily concerns annotation efficiency, applicability at the network scale, and structural stability under complex observational conditions rather than the underlying cloud-detection principle itself.

We also removed the potentially misleading term “calibration-decoupled.” The model does not use absolutely calibrated backscatter or extinction coefficients as inputs and therefore reduces its dependence on absolute energy calibration and assumptions involved in optical retrieval. However, the input signals still undergo background and range corrections, and the VDR requires polarization gain-ratio calibration. The revised manuscript therefore consistently uses the expression “reduced dependence on absolute calibration” and explicitly states that the method is not completely independent of calibration. Correspondingly, the demonstrated transferability is limited to the three held-out CARLNET sites with similar observation channels and processing protocols. It should not be extrapolated to universal transferability across different lidar systems.

Location of revision

Abstract, Lines 15–31; Introduction, Lines 94–112; Section 2.2, Lines 168–174; Conclusion and Discussion, Lines 658–664 and 692–705.

Relevant revised text

“The methodological contribution of this study lies in the lidar-network-specific integration of self-supervised representation learning, empirical VDE-based probabilistic guidance, and limited expert supervision into a unified workflow for cross-site Raman lidar observations.”

“This study developed and validated a multi-channel cloud-layer identification framework with reduced dependence on absolute calibration for automated Raman lidar network processing.”

Major comment 2 Dataset and split strategy

Reviewer comment

2. The dataset description and split strategy require clarification

The description of the CARLNET dataset needs to be more precise. The statement that the dataset was expanded from 47 sites to 79 sites “according to personal communication” may be not appropriate. The reported dataset size (“approximately 15,000 days”) is also ambiguous. It should be clarified whether this refers to site-days, operational days, lidar measurement days, or another quantity. The training, validation, and testing strategy should be described more explicitly. The manuscript indicates that

44 sites are used for training/validation and three sites are reserved for independent testing, while an approximately 80/20 training-validation split is also mentioned. The authors should clarify: whether validation samples are drawn from the same 44 training sites; whether the validation split is random, temporal, or site-based; how data leakage between training and validation is prevented. These details are essential for reproducibility.

Author response

We removed the statement referring to 79 sites and “personal communication.” We now clarify that this study uses 47 CARLNET sites that met the quality-control requirements and that the dataset comprises approximately 15,000 site-days, rather than 15,000 calendar days or operational days from a single site.

Regarding the splitting strategy, 44 sites were used as development sites, and the validation samples were indeed drawn from these same 44 sites rather than from a separate set of validation sites. The training-validation split was not performed randomly at the level of individual, highly overlapping 10-profile windows. Instead, it was performed at the site-day level: approximately 80% of the site-days were assigned to training and 20% to validation. All 10-profile samples constructed from the same site on the same day were assigned exclusively to one subset. This prevents adjacent and strongly overlapping temporal windows from the same day from being shared between the training and validation subsets.

Data from different dates may nevertheless originate from the same development site. Therefore, the validation subset is used primarily for early stopping and for hyperparameter and threshold selection; it is not interpreted as an independent cross-site generalization test. The actual cross-site and cross-time evaluation is performed using the three completely held-out test sites. Their continuous observations from January 2025 were excluded from pretraining, fine-tuning, threshold selection, hyperparameter selection, and early stopping.

Location of revision

Section 2.1, Lines 120–149; Section 2.4, Lines 244–251; Appendix A7, Lines 845–850.

Relevant revised text

“Within the 44 development sites, the training-validation split was performed at the site-day level, with approximately 80% of the site-days assigned to training and 20% assigned to validation. All 10-profile samples constructed from the same

site on the same day were assigned exclusively to either the training or validation subset, thereby preventing highly overlapping temporal samples from being shared between the two subsets.”

“The independent test set [...] was excluded from model training, hyperparameter selection, threshold selection, and early stopping.”

Major comment 3 Methodological choices

Reviewer comment

3. Several methodological choices require stronger justification

The preprocessing pipeline is reasonable and physically motivated, including background correction, range correction, normalization, logarithmic transformation, and height-dependent compression. The anisotropic treatment of temporal and vertical dimensions is also appropriate for lidar observations. However, several key parameters appear empirically selected and should be supported by either sensitivity analyses or clearer physical justification. Why are 10 consecutive profiles (50 min) selected? How does performance change with shorter or longer windows, particularly for rapidly evolving clouds? Why are the boundaries selected at 0–12 km, 12–24 km, and 24–30 km? What physical considerations motivate these thresholds? Why is 0.65 chosen as the final classification threshold? How sensitive are cloud-base and cloud-top errors to this choice? Why is a stride of 10 used during training but 1 during inference? How does this affect the effective training sample size and model generalization? These design choices should be supported by experiments or a concise rationale.

Author response

We thank the reviewer for this suggestion. We added the rationale for each of these parameter choices in the main text and Appendix while avoiding any implication that these empirical settings are universally optimal.

First, each sample consists of 10 consecutive profiles with a temporal resolution of 5 min, corresponding to a local temporal window of 50 min. This setting allows the model to exploit short-term cloud continuity and multi-channel structural information without depending on long, uninterrupted observation sequences. A 50 min window covers the timescale of many cloud-layer structural variations while reducing the influence of operational interruptions and missing observations on sample construction. The effect of different window lengths on rapidly evolving clouds warrants a dedicated investigation. Because no systematic window-length sensitivity analysis was conducted

in this study, the 10-profile setting is described only as a physical and operational compromise rather than as a theoretically optimal value.

Second, the 0–12, 12–24, and 24–30 km segmentation is designed to accommodate the different information densities and signal characteristics of the three altitude ranges. The 0–12 km region contains most tropospheric clouds and the cloud bases and boundaries of primary operational interest; relatively fine vertical sampling is therefore retained in this region. At 12–24 and 24–30 km, the frequency of cloud occurrence and the amount of effective echo information gradually decrease, while the redundancy in vertical sampling increases. Resampling factors of 4 and 8 are therefore applied, respectively, to reduce computational cost and enlarge the effective representation scale for weak high-altitude structures. These boundaries represent an empirical compromise among boundary preservation, signal characteristics, and computational efficiency rather than strict physical atmospheric-layer boundaries.

Third, the final threshold of 0.65 was selected using the held-out validation subset from the 44 development sites to balance missed detections and false positives. It was fixed before evaluation at the three test sites, and no test-site data were used for threshold selection. The manuscript does not report complete sensitivity curves for cloud-base and detectable cloud-top errors under different thresholds. The revised manuscript therefore explicitly limits the interpretation of 0.65 to an operational threshold under the current dataset split and model configuration rather than a universal threshold applicable to all lidar systems.

Finally, a stride of 10 is used during training to reduce the strong overlap between adjacent 50 min windows, avoid overcounting large numbers of nearly duplicated samples, and reduce computational redundancy. This decreases the nominal number of training samples but increases their relative independence. During inference, a stride of 1 is used to generate predictions at every time step and produce continuous cloud masks without temporal gaps. Because no independent ablation of different stride settings was conducted, their effects are explained in terms of controlling data redundancy and ensuring continuous inference, without claiming that the selected settings provide optimal generalization performance.

Location of revision

Section 3.1, Lines 330–335; Section 3.2, Lines 343–352; Appendix A1, Lines 719–742; Appendix A6, Lines 823–834.

Relevant revised text

“Each sample consists of 10 consecutive lidar profiles with a temporal resolution of 5 min, corresponding to a 50-min local observation window, which is comparable to the timescale of many cloud-layer structural variations while remaining robust to operational interruptions and missing observations.”

“After Stage 3, the final operational threshold of 0.65 is selected using the held-out validation subset to balance missed detections and false positives.”

Major comment 4 Label generation and annotation bias

Reviewer comment

4. The label-generation process and potential annotation bias require further discussion

The auxiliary products influence the training labels, the potential propagation of retrieval uncertainties into the labels should be discussed more explicitly. This issue is particularly relevant because Fernald-based retrievals are used when Raman retrievals are unavailable or unstable. Since Fernald retrievals involve assumptions and may introduce systematic uncertainties, the authors should discuss how such uncertainties may affect the final model. Furthermore, although two experts independently annotated the data, no quantitative measure of annotation consistency is provided. Reporting inter-annotator agreement metrics (e.g., Cohen's kappa, pixel-wise agreement, or IoU) for a representative subset would substantially strengthen the reproducibility and reliability of the training labels.

Author response

We thank the reviewer for this suggestion. We recognize that the wording in the original manuscript may have implied that the entire manually labeled dataset was independently annotated in full by two experts. The actual procedure was as follows. The primary manual labels were produced by experts with experience in lidar operation and research. To quantitatively assess inter-annotator consistency, a representative subset of 15 site-days, corresponding to approximately 10 % of the manually labeled dataset, was independently annotated by two experts without access to each other's labels. The two annotation sets achieved a pixel-wise IoU of 95.8 %. Regions of disagreement were subsequently reviewed jointly, and the final labels were determined through expert discussion and consensus.

We have revised the manuscript to distinguish the overall label-generation procedure

from the independent consistency assessment performed on this representative subset. We also clarify that the reported IoU characterizes inter-annotator consistency within the examined subset and does not constitute a complete quantification of all labeling uncertainty or potential bias shared by the experts.

Location of revision

Section 2.3, Lines 184–217; Section 2.4, Lines 219–247.

Relevant revised text

“Manual annotation was performed by experts with experience in lidar operation and research. To quantitatively assess inter-annotator consistency, a representative subset of 15 site-days, corresponding to approximately 10 % of the manually labeled dataset, was independently annotated by two experts without access to each other’s labels. The two annotation sets achieved a pixel-wise IoU of 95.8 %. Disagreements were subsequently reviewed and resolved through expert discussion and consensus.”

Major comment 5 Potential inheritance of VDE bias

Reviewer comment

5. The potential inheritance of VDE bias should be discussed more explicitly

Although VDE is not used during final inference, the student model may still inherit some systematic biases from the teacher during distillation. The manuscript already notes that VDE can produce enhanced cloud probabilities under strong aerosol loading. This limitation should be discussed as an inherent trade-off: VDE improves label efficiency and introduces physical constraints, but it may also transfer empirical-rule limitations into the learned representation. A clearer discussion of this balance would improve the interpretation of the proposed framework.

Author response

We agree that excluding VDE from the final inference procedure does not imply that its biases cannot be retained by the student model. Stage 2 uses VDE-derived probabilities as weak supervision. Consequently, empirical-rule biases, such as elevated cloud probabilities under strong aerosol loading, may be partially encoded in the learned representations. The revised manuscript explicitly describes this as an inherent trade-off

of VDE-guided distillation: it provides structural priors and probabilistic supervision when manual annotations are limited, but it may also transfer systematic limitations of the teacher rules.

We explain the mitigation of this influence from three perspectives. First, Stage 2 uses soft probabilistic targets rather than directly reproducing the binary VDE output. It also incorporates multi-wavelength information, aerosol-suppression constraints in regions with low teacher cloud scores, and near-range constraints. Second, Stage 3 further optimizes the decision boundary using expert-refined annotations, allowing the final model to correct some VDE patterns that are inconsistent with expert interpretation. VDE is also not used during inference. Third, the final model produces results that are clearly different from those of VDE. It exhibits different boundary structures in terms of cloud detection rate, cloud-base and detectable cloud-top errors, and complex aerosol–cloud coexistence cases. Across the 47 matched radiosonde cases, its structural consistency rate is 38.3%, compared with 34.0% for VDE. These results indicate that the student model does not simply reproduce the VDE rules but instead combines, to some extent, the structural priors provided by VDE with the boundary information contained in the expert annotations.

Nevertheless, these differences demonstrate only that the influence of VDE bias is mitigated; they do not demonstrate that it has been completely eliminated. Nor can the radiosonde-based consistency results be interpreted as proof of absolute physical correctness. The revised manuscript therefore retains the limitation that VDE bias may be partially inherited.

Location of revision

Section 2.6, Lines 275–304; Section 3.3, Lines 387–399; Sections 4.2–4.3, Lines 482–640.

Relevant revised text

“Because the VDE-derived probabilities are used as weak supervision during Stage 2, systematic errors in the empirical teacher, such as elevated cloud probabilities under strong aerosol loading, may be partially inherited by the student model. The use of soft probabilistic targets, multi-wavelength averaging, explicit aerosol-suppression constraints, and subsequent fine-tuning with expert annotations is intended to reduce this influence, but cannot eliminate it completely.”

Major comment 6 Radiosonde as a consistency check

Reviewer comment

6. Radiosonde comparison should be presented strictly as a consistency check

The radiosonde comparison is valuable, but its role should be more clearly defined. Radiosonde-derived cloud boundaries are affected by threshold selection, vertical sampling, humidity criteria, and horizontal drift during ascent; therefore, they should not be interpreted as ground truth. The manuscript appropriately acknowledges these limitations, but some statements may imply that radiosonde agreement directly validates physical correctness. The authors should emphasize that radiosonde comparisons provide an independent consistency assessment at the structural or trend level, rather than a strict validation benchmark.

Author response

We fully agree with this comment and have redefined the evidential role of the radiosonde comparison. This study uses 47 matched lidar–radiosonde cases from the three held-out test sites, rather than observations from 47 radiosonde sites. The relative-humidity thresholding method identifies moist layers sampled along the drifting balloon trajectory, whereas the lidar identifies scattering layers above a fixed site. The two systems therefore differ in the observed physical quantities, temporal and spatial sampling, and boundary definitions. The radiosonde results are also affected by the humidity thresholds, vertical sampling, moist-layer merging rules, and horizontal drift during ascent. Consequently, radiosonde observations cannot be treated as pixel-level or boundary-level ground truth.

In the revised manuscript, the corresponding metric is consistently termed the “structural consistency rate” and is used only to assess agreement in layer altitude, number of layers, continuity, and the principal evolution patterns. The differences among 38.3%, 34.0%, and 27.7% are not interpreted as differences in the absolute accuracy of the three methods and cannot directly demonstrate that the model has greater physical correctness. Instead, the radiosonde observations provide a reference that is independent of the manual lidar annotations and allow us to examine whether the principal structural conclusions are broadly consistent with another type of atmospheric observation. The wording in the Abstract, results, figure captions, and conclusions was moderated according to this level of evidence.

Location of revision

Section 2.5, Lines 270–272; Section 4.3, Lines 622–640; Figure 7 caption, Lines 644–647; Abstract, Lines 25–29; Conclusion and Discussion, Lines 681–687.

Relevant revised text

“Because radiosonde-based identification is affected by humidity thresholds, ascent sampling, and spatial drift, these statistics are interpreted as structural consistency indicators rather than strict accuracy measurements.”

Major comment 7 Uncertainty analysis

Reviewer comment

7. Performance improvements require stronger uncertainty analysis

The pretrained model improves F1 from 0.9281 to 0.9371 and reduces cloud-base and cloud-top MAE compared with the baseline. These results indicate the value of the proposed framework. However, the statistical robustness of these improvements remains unclear. The manuscript does not report variability among repeated experiments with different random seeds, confidence intervals for key metrics and statistical significance tests for error reductions. Uncertainty estimates are needed to determine whether the reported gains are stable across different training conditions.

Author response

We agree with the reviewer. Repeated training with different random seeds, confidence intervals, and statistical tests can quantify uncertainty arising from model initialization and optimization and are effective approaches for evaluating the stability of the reported performance gains.

The present study primarily focuses on the operational applicability of a fixed training strategy under strict cross-site and cross-time conditions. All models use the same dataset split, training configuration, and evaluation procedure and are tested using continuous observations from three completely held-out CARLNET sites. The current results show that the benefits of pretraining are not restricted to a single F1 metric. The F1 score increases from 0.9281 to 0.9371; the detectable cloud-top MAE decreases from 94.0 to 82.2 m, while the cloud-base MAE remains comparable. The complex cases in Fig. 6, the cloud detection rates, and the independent radiosonde-based structural consistency assessment also show directionally consistent results. These complementary findings suggest that the main contribution of pretraining lies in the

more stable representation of weak cloud layers and complex cloud structures rather than simply increasing the number of detected cloud layers.

Nevertheless, we agree that these results cannot replace experiments with multiple random seeds and cannot be used to claim that the performance differences are statistically significant. Multiple-seed experiments mainly characterize uncertainty in the training process, whereas operational network applications must also consider sampling uncertainty associated with different sites, temporal windows, and seasonal conditions. A comprehensive uncertainty assessment of this framework would require repeated training over longer time series, together with site-wise or site-day-wise statistics and confidence-interval estimation. Adding a limited number of random-seed experiments for only a small subset of configurations during the present revision would still be insufficient to address all of these issues comprehensively.

We therefore adopted a more transparent and conservative approach in this revision. We now explicitly state that Figs. 4–6 and Tables 2–3 were obtained from a single training run under a fixed configuration. The F1 improvement is described as a “modest deterministic performance gain” rather than as a statistically significant improvement. The Conclusion also identifies systematic uncertainty assessment based on longer time series, different random initializations, and confidence intervals as an important direction for future work.

Location of revision

Section 4, Lines 449–451; Section 4.2.1, Lines 491–500; Conclusion and Discussion, Lines 700–705.

Relevant revised text

“The results reported in Figs. 4–6 and Tables 2–3 were obtained from a single training run using the fixed experimental configuration described in Appendix A. They are therefore interpreted as deterministic performance estimates for the specified data split rather than as statistical estimates of variability across model initializations.”

Minor comment 1 Terminology and wording

Reviewer comment

1. Terminology and wording

“Calibration-decoupled” should be revised to “reduces dependence on absolute

calibration” or similar, more precise language.

“According to personal communication” should be removed or replaced with a verifiable citation.

“Approximately 15,000 days” should be clarified with the correct unit (e.g., site-days or total lidar operational days).

The wording throughout should be precise and reproducible.

Author response

We revised each of these expressions. The term “calibration-decoupled” was replaced throughout the manuscript with “reduced dependence on absolute calibration,” reflecting the fact that the model still relies on basic signal processing and polarization gain-ratio calibration for VDR. The unverifiable reference to “personal communication” and the statement concerning 79 sites were removed. The dataset size is now specified as approximately 15,000 site-days. Here, a site-day represents observations from one site on one measurement day, thereby avoiding confusion with calendar days or the number of consecutive operational days at a single site.

Location of revision

Abstract, Lines 15–31; Introduction, Lines 94–112; Section 2.1, Lines 120–130; Section 2.4, Lines 244–251; Conclusion and Discussion, Lines 658–664.

Minor comment 2 Figures and tables

Reviewer comment

2. Figures and tables

For Figure 1, explicitly state the criteria for selecting the three test sites (e.g., to maximize climatic/environmental diversity).

For Figures 4–6, if the results are from a single run, this should be stated. If they are averages, error bars or confidence bands would be a significant enhancement.

For Tables 2 and 3, include uncertainty information (e.g., standard deviations over multiple runs or bootstrap confidence intervals) to provide statistical context for the reported differences.

For Figures 7 and 8, reinforce the caption to clearly state that radiosonde is used as a trend-level consistency reference, not a ground-truth benchmark.

Author response

We added the criteria for selecting the three test sites in Section 2.1. The three sites represent an arid inland environment, a monsoon-influenced region, and a humid coastal environment, respectively. They also differ in aerosol background, cloud structure, and observational conditions and therefore provide an environmentally diverse independent cross-site test.

Figs. 4–6 and Tables 2–3 do not present averages over repeated training runs; they report results from a single run under a fixed dataset split and training configuration. Because no repeated samples across random seeds were available, we did not add potentially misleading error bars, standard deviations, or confidence bands. Instead, Section 4 explicitly identifies these results as deterministic performance estimates, and the corresponding statistical limitation is acknowledged in the Conclusion. The radiosonde comparison is consistently presented as a structural and trend-level consistency reference rather than as a ground-truth benchmark.

Location of revision

Section 2.1, Lines 137–153; Section 4, Lines 449–451; Section 4.3, Lines 622–640; Figure 7 caption, Lines 644–647.

Minor comment 3 Moderation of conclusions

Reviewer comment

3. Several conclusions should be moderated

Some statements in the abstract and conclusions appear stronger than supported by the current evidence. For example, “Calibration-decoupled” should be weakened as discussed above. “Transferable” and “cross-site deployment”: the study demonstrates generalization to three held-out CARLNET sites, but broader transferability across different lidar systems, wavelengths, and geographic regions remains untested. “Improved physical plausibility”: the results demonstrate improved continuity and coherence of cloud masks, but physical plausibility is not directly quantified.

The conclusions should better reflect the demonstrated capabilities and acknowledged limitations, including blind-zone effects, daytime background noise, saturation, and strong attenuation.

Author response

We agree that some statements in the original manuscript extended beyond the available evidence and have revised them according to the level of support provided by the results. “Calibration-decoupled” was replaced with “reduced dependence on absolute calibration.” The terms “transferable” and “cross-site deployment” are now restricted to cross-site generalization across the three held-out CARLNET sites within the same network observation system. They no longer imply that universal transferability has been demonstrated across other wavelength configurations, instrument types, or geographic regions. The radiosonde results are also no longer used to claim direct validation of “physical correctness” but only to support consistency in cloud-layer altitude, continuity, and multilayer structure.

In addition, the Conclusion explicitly retains the limitations arising from incomplete near-range overlap, daytime background noise, signal saturation, strong attenuation and complete obscuration by optically thick low clouds, irregular temporal sampling, and prolonged data gaps. In particular, when the lidar signal is completely attenuated below the physical cloud top, the model can provide only the detectable cloud top and cannot recover the true physical cloud top. Cross-instrument generalization to external lidar networks also requires further evaluation.

Location of revision

Abstract, Lines 24–31; Section 4.3, Lines 622–640; Conclusion and Discussion, Lines 658–705.

Relevant revised text

“Although the current validation uses strict held-out sites within CARLNET, broader tests on external lidar networks are needed to assess cross-instrument and cross-platform generalizability. In addition, the current evaluation focuses on deterministic model performance under a fixed experimental setting.”

Minor comment 4 References

Reviewer comment

4. References

The reference list is adequate overall, but the manuscript would be strengthened by more explicit citation of literature on lidar network transferability, uncertainty in cloud annotation from lidar, and the known limitations of Fernald-based retrievals used for labeling.

Author response

Following this suggestion, we strengthened the literature support directly relevant to the proposed methodology without expanding the discussion into a general review of lidar-network transferability. Specifically, Shao et al. (2025) is cited to describe the CARLNET observational network and data-processing background. Klett (1981) and Fernald (1984) are cited to explain the sensitivity of elastic-signal retrievals to the assumed lidar ratio and reference-region selection. Mylonaki et al. (2021), Floutsi et al. (2023), and related studies support the discussion of cloud–aerosol discrimination, attenuation, and uncertainty in the auxiliary products. The corresponding text also clarifies that these retrieved products are used only to support expert interpretation and are not used as model inputs or deterministic label-generation criteria.

Location of revision

Sections 2.1–2.4, Lines 120–251; References.

Relevant revised text

“The resulting elastic-signal-based estimates remain sensitive to the assumed lidar ratio and reference-region selection (Klett, 1981; Fernald, 1984) and are therefore used only as auxiliary information for expert interpretation rather than as deterministic annotation criteria.”