# Response to Reviewers' Comments

## Reviewer 1

### Overall evaluation

**Reviewer comment (in blue)**

The manuscript proposes a self-supervised and AI framework for cloud-layer identification in multi-wavelength Raman-polarization lidar networks. The work addresses an important gap in operational lidar processing. Therefore, the topic is generally very well suited for AMT. By operating directly on uncalibrated elastic backscatter and calibrated volume depolarization ratio channels from the Chinese lidar network, the approach aims to reduce dependence on absolute calibration and improve robustness across heterogeneous network sites with the potential to be applied to other lidar networks as well.
The three-stage training pipeline—self-supervised reconstruction pretraining, VDE-guided probabilistic distillation, and fine-tuning with limited expert labels—is to my opinion a complete novel and very sophisticated approach. It seems solid, and the reported gains are promising. Therefore, well suited for publication in this journal for the atmospheric research community.
However, for a good understanding, the manuscript needs additional methodological description (more details) and clearer justification of certain design choices (especially the descriptions given in the Annex). Thus, while generally well written, the methodological section needs to be strengthened and therefore recommend major revisions before the paper can be accepted for publication.

**Author response (in black)**

We thank the reviewer for the positive assessment of the significance, overall framework, and potential applications of this study. We understand that the reviewer's primary concern is not the soundness of the framework itself, but that the original manuscript did not sufficiently explain its methodological logic, key design choices, and details required for reproducibility. We therefore systematically revised the methodology and experimental sections. Specifically, we expanded the descriptions of the CARLNET data sources, input channels, and dataset split; explained the anisotropic encoder–decoder architecture and vertical resampling procedure; clarified the respective roles and connections of self-supervised representation learning, VDE-guided probabilistic distillation, and expert-supervised fine-tuning in the three-stage training workflow; and provided additional details on the loss functions, masking ratios, augmentation parameters, thresholds, early-stopping criterion, and inference settings. We

also expanded the evaluation procedures for pixel-level metrics, cloud-base and detectable cloud-top heights, and radiosonde-based structural consistency.

In addition, we carefully reconsidered the statements concerning methodological novelty and generalizability. The contribution of this study does not lie in introducing a new physical mechanism of cloud scattering, but in integrating structural learning from unlabeled observations, probabilistic weak supervision provided by a conventional algorithm, and limited expert supervision into a unified workflow designed for multiwavelength polarization Raman lidar networks. Accordingly, we now describe the method as “reducing dependence on absolute calibration” and restrict the demonstrated transferability to the three completely held-out CARLNET test sites.

**Location of revision**

*Revisions were made throughout Sections 2–4 and Appendix A, particularly Section 3.3 (Lines 379–408) and Appendix A1–A7 (Lines 719–854).*

## General comment 1 Repetition and methodological balance

**Reviewer comment**

Sometimes the manuscript is very repeating and thus longish to read in the current form. Please try to sharpen. E.g. I have read 10 times that “the 3 test sites are not used for training”. On the other hand—as described above—some methodological descriptions are too short (partly specified in the specific comments). Please try to find a good balance.

**Author response**

We removed or consolidated repetitive statements and concentrated the rules governing the use of the test sites in the dataset-splitting and experimental-setting sections. Meanwhile, we expanded methodological details that were previously described too briefly. A small amount of necessary repetition was retained to clarify the dataset split, main experimental protocol, and interpretation of the results in their respective contexts.

**Location of revision**

*Section 2.1, Lines 137–149; Section 4, Lines 438–451; Section 4.2.1, Lines 482–510; Appendix A, Lines 714–860.*

## General comment 2 Terminology and wording

### Reviewer comment

The English language level is very good, probably AI supported—which is ok, even though there were many vocabularies I've never heard before and I had to check out. But please check if all wordings can be used in the context. E.g., distilling/distillation, cue, pervasive, ablation. I am not sure if distilling is the correct word for what you want to say even though I understand the meaning. I guess only a native speaker could help there.

### Author response

We re-examined and standardized the relevant terminology throughout the manuscript. The terms "knowledge distillation" and "probabilistic distillation," which are commonly used in machine learning, were retained, and their meanings are now explained at their first occurrence. Potentially ambiguous expressions, such as "adaptation ablation," were removed or replaced. We also revised the wording to make the manuscript more accessible to readers in atmospheric remote sensing.

### Location of revision

*Introduction, Lines 94–112; Section 3.3, Lines 379–408; Section 4, Lines 438–447.*

#### Relevant revised text

*"Here, self-supervised learning is used to learn transferable structural representations from unlabeled lidar observations, while knowledge distillation introduces probabilistic guidance from a conventional cloud-detection method as weak supervision." "Specifically, discretized cloud-probability levels generated by the conventional VDE algorithm [...] are used as probabilistic supervisory information rather than absolute ground truth labels [...]."*

## General comment 3 Explanation of AI terminology

### Reviewer comment

Furthermore, it would be great if some of the "AI slang" could be explained in a more detailed way as this is a manuscript in an atmospheric research journal. E.g., U-Net is introduced without any further description.

### Author response

We added functional explanations of U-Net and its associated modules rather than merely naming the architecture. The revised manuscript explains that the encoder progressively extracts multiscale structural features from the time–height input; the decoder gradually restores spatial resolution to generate a pixel-level cloud mask; and the skip connections transfer fine-scale cloud-boundary information from shallow layers to the decoding stage, thereby reducing the loss of boundary details during downsampling.

The Appendix further explains several terms that may be unfamiliar to readers. The segmentation head is the output module that converts network features into a single-channel cloud/non-cloud probability map. The reconstruction head is used only in Stage 1 to recover the masked input signals. BCE-with-logits combines the sigmoid operation and binary cross-entropy calculation in a numerically stable formulation for probabilistic segmentation training. These additions allow readers without a specialized background in artificial intelligence to understand the functional role and training objective of each module.

**Location of revision**

*Section 3.2, Lines 337–377; Appendix A3–A4, Lines 762–806.*

**Relevant revised text**

*"In this framework, the encoder extracts hierarchical representations from the time–height input, while the decoder progressively restores spatial resolution to generate pixel-level cloud-layer masks."*

## General comment 4 Methodological clarity

**Reviewer comment**

Generally, sometimes methodological clarity needs to be further sharpened. Most descriptions are top level and more detail is needed.

E.g.:

1. The self-supervised pretraining tasks (channel-wise masking, temporal frame prediction) are described only at a high level. Precise formulation of loss functions, masking ratios, and the balance between reconstruction and prediction objectives is needed.

2. Reported metrics include IoU, F1 score, detection rates, …. Partly the reader should be more guided for the interpretation of these metrics—description is a bit missing and should be extended.

3. Numbers used for the approach like penalties are given mostly without any justification. This should be improved. If the choice was based on empirical studies, it is ok, but should be stated.

4. Cloud top: It remains unclear for me how cloud tops are defined. As there might be many cases where signal attenuation is strong and a cloud top height cannot be detected with lidar, I wonder how these circumstances are treated. Under which conditions are you able to identify cloud top heights properly?

5. Radiosonde comparison: It is not clear if radiosondes are available at all sites and have been used for evaluation or only on the selected ones. Please clarify.

6. Temporal context window: The fixed 10-profile window: It remains unclear if it is 5 min or 50 min in total. Please clarify.

7. In the evaluation of the proposed methodology, models trained with and without pretraining are compared. It is unclear which parts of the 3-stage training process are included in the pretraining. Only Stage 1—reconstruction, or Stage 1 and Stage 2—reconstruction and distillation. Please clarify this.

**Author response**

We thank the reviewer for identifying these points requiring clarification. We addressed the seven questions individually, as follows.

1. For self-supervised pretraining, Appendix A4 now provides specific definitions of the channel-masking, temporal-frame prediction, and full-signal reconstruction tasks, together with the corresponding loss functions, masking ratios, and loss weights. The purpose of Stage 1 is not to identify clouds directly, but to learn cross-channel relationships, short-term temporal continuity, and vertical structures from unlabeled observations.
2. Appendix A7 now provides the definitions and interpretations of precision, recall, F1, and IoU. The F1 score summarizes the balance between missed detections and false alarms, whereas IoU provides a stricter measure of the spatial overlap between the predicted and reference cloud masks. Cloud-detection rate and geometric errors provide additional product-oriented assessments of whether cloud layers are detected and whether their bases and detectable tops are correctly located.
3. We now clarify that parameters such as augmentation magnitudes, loss weights, penalty terms, and probability thresholds were selected empirically using observations from the development sites and the held-out validation subset. Data

from the test sites were not involved in parameter selection. Because we did not perform an exhaustive search over all parameter combinations, these values are not presented as universally optimal settings.

4. We strictly define the cloud top as the “detectable cloud top,” namely, the highest cloud boundary supported by valid lidar returns. During manual annotation, a mean SNR of 3 across the three elastic channels is used as the threshold delimiting the valid signal range. When optically thick clouds cause strong attenuation such that the mean SNR above the cloud remains below 3, the annotation extends only to the highest altitude supported by a valid return. This truncation altitude is not interpreted as the physical cloud top. Because the true cloud top is unobservable under complete attenuation, such cases are excluded from the cloud-top boundary-error evaluation. Detectable cloud-top errors are calculated only when sufficient return signals remain near and above the cloud top.

   We emphasize that the SNR threshold of 3 is used only to determine the valid signal range during manual annotation and is not applied as a hard decision threshold during model inference. In some strongly attenuating cases, the model produces more spatially continuous cloud structures than the directly supervised model. However, because no independent observations are available within the fully attenuated region, we do not interpret this continuity as accurate recovery of the unobservable physical cloud top. In operational deployment, an SNR-based signal-validity mask or quality flag could be introduced, where necessary, to restrict model outputs in regions without valid returns. The appropriate threshold would nevertheless require further assessment using longer observation periods and a larger number of attenuation cases.

5. Co-located operational radiosonde observations are available at all 47 CARLNET sites, but the present study does not use radiosonde data from all sites. To maintain consistency with the strict cross-site testing protocol, the radiosonde comparison uses only observations from the three independent test sites during the continuous test period in January 2025. After temporal matching and quality control, 47 valid lidar–radiosonde matched cases were retained, and the corresponding statistics are summarized in Table 4. Radiosonde observations were not used for model training, manual annotation, hyperparameter selection, or threshold determination. They were used only for the final independent structural-consistency assessment and were not treated as strict ground truth. We have clarified the overall availability and actual use of the radiosonde data in Sections 2.5 and 4.3.

6. Each sample contains 10 consecutive profiles, with a temporal resolution of 5 min per profile. The total temporal context window is therefore 50 min rather than 5 min.
7. The revised manuscript explicitly defines “pretraining” as the combination of Stages 1 and 2. Stage 1 performs self-supervised representation learning, while Stage 2 performs VDE-guided probabilistic distillation. Together, they constitute the physics-guided pretraining phase, which is followed by supervised fine-tuning in Stage 3 using a limited number of expert-refined labels.

These additions clarify the logic of the three-stage training workflow. Stage 1 first learns stable structural representations from a large volume of unlabeled observations. Stage 2 then uses physically relevant weak supervision from the conventional algorithm to constrain the decision boundary. Finally, Stage 3 uses high-quality expert annotations to correct potential biases inherited from the teacher algorithm.

**Location of revision**

*Section 3.1, Lines 322–335; Section 3.3, Lines 379–408; Section 4.2.2, Lines 512–535; Section 4.3, Lines 598–605; Appendix A4–A7, Lines 771–854.*

**Relevant revised text**

*“In this study, the term ‘pretraining’ refers to the first two stages, including self-supervised representation learning (Stage 1) and VDE-guided probabilistic distillation (Stage 2). Together, these two stages constitute a physics-guided pretraining phase before supervised fine-tuning with expert annotations (Stage 3).”*

## General comment 5 Figure labels and self-contained captions

**Reviewer comment**

Labels of figures partly need to be extended. As a general rule, figures including their labels should be as self-explaining as possible. E.g. in Fig. 3, it is not described what the numbers in brackets mean and just by reading all text one can guess what is meant. See further specific comments below. Please change this.

**Author response**

The caption of Fig. 3 has been substantially expanded. We re-examined the symbols, tensor dimensions, line styles, and panel labels in all figures. For Fig. 3, we standardized the vertical resampling ranges and factors across the main text, Appendix, and figure. We also clarify that the input dimensions “[5, 10, 1056]” represent five observational channels, 10

consecutive temporal profiles, and 1056 resampled vertical bins, respectively. In addition, we distinguish the vertical resampling applied before the input enters the network from feature downsampling within the network architecture, thereby avoiding the interpretation of both operations as “pooling.”

The captions of Figs. 5–8 have also been expanded to explain the methods, statistics, line styles, boundary definitions, and comparison regions represented in the individual panels. These revisions allow readers to understand the principal information presented in each figure without repeatedly referring to the main text.

**Location of revision**

*Section 3.2, Lines 337–377; Figure 3; Figures 5–8, Lines 559–564, 593–596, 644–656.*

## General comment 6 CARLNET data sharing and international coordination

**Reviewer comment**

The CARLNET network looks impressive. In general, it would be great if CARLNET data could be shared in future and the initiative joins activities within GAW (GALION) and/or contributes to other global activities like satellite validation.

**Author response**

We thank the reviewer for this suggestion. The Data Availability statement now clarifies that access to the complete CARLNET observational dataset is subject to institutional data-management policies. Standardized data products will be released in the future where permitted by these policies. Cross-validation with international lidar networks and satellite observations is also an important direction for future work.

**Location of revision**

*Code and data availability, Lines 862–866; Conclusion and Discussion, Lines 706–711.*

## General comment 7 Code and training-data availability

**Reviewer comment**

The approach presented seems to have great potential to be checked against other lidar networks. Therefore, I would insist to share the code and training data as promised during the publication process. For a peer-review setting, providing a DOI of the code would be key.

**Author response**

We have clarified in the availability statement that the training and inference code, pretrained model weights, and transferable application examples will be made publicly available after publication (which is hosted privately in GitHub, https://github.com/WHU-Atmospheric-Remote-Sensing/Cloud_detection_lidar). The data and scripts required to reproduce the main figures and tables can be provided during peer review upon request. Release of the complete observational dataset remains subject to institutional data policies. A permanent repository link or DOI will be added once the repository is formally released.

**Location of revision**

*Code and data availability, Lines 862–866.*

## Specific comment 1 Abstract F1 improvement

**Reviewer comment**

Abstract – The reported F1 improvement (0.9281 → 0.9371) is presented without context and statement if it is significant. For my understanding, the improvement is negligible. Please clarify.

**Author response**

We agree with this comment. In the original Abstract, the values 0.9281 and 0.9371 were presented directly alongside each other, which could overemphasize the relatively small absolute difference without clarifying its statistical meaning. The revised Abstract no longer highlights the four-decimal difference. Instead, it reports the final F1 score—a metric that balances missed detections and false positives—together with the more operationally interpretable errors in cloud-base and detectable cloud-top heights.

In the main text, the difference is explicitly described as a “modest deterministic performance gain” under a fixed dataset split and a single training configuration, rather than as evidence of statistical significance. Expressed in terms of the error proportion, $(1 - F1)$, the change corresponds to a relative reduction of approximately 12.5%. Its operational relevance is supported primarily by complementary results: the pretrained model provides cloud-base accuracy comparable to that of the directly supervised model, while reducing the detectable

cloud-top MAE from 94.0 to 82.2 m; it also produces more continuous weak-cloud and multilayer structures in complex cases. We therefore interpret the F1 result as one component of the overall evidence rather than using it alone to claim a statistically significant improvement.

**Location of revision**

*Abstract, Lines 15–31; Section 4.2.1, Lines 491–500; Conclusion and Discussion, Lines 665–680.*

**Relevant revised text**

*"Because the results are obtained under a fixed dataset split and training configuration, the F1 difference should be interpreted as a modest deterministic performance gain rather than formal evidence of statistical significance."*

## Specific comment 2 Figure 1 inset and dashed lines

**Reviewer comment**

Figure 1: The small embedded map in the lower right is not needed and rather misleading. I would recommend removing it. In general, the dashed lines are not explained; if needed, please do so—otherwise remove.

**Author response**

We re-examined Fig. 1. The lower-right inset is retained because it forms part of the cartographic elements required for the complete representation of a map of China under the current national mapping specifications. Specifically, Sections 5 and 6 of the *Specifications for the Representation of Public Map Content* (Natural Resources Regulation [2023] No. 2) state that a complete map of China should include the South China Sea islands, which may be presented as an inset. The official document is available at the link below:

https://www.mfa.gov.cn/web/wjb_673085/zzjg_673183/bjhysws_674671/bhflfg/dtdmxgfl/202303/P020230313585504979937.pdf?utm_source=chatgpt.com

The inset is retained solely to preserve the geographic completeness of the base map and is not used for site selection, dataset partitioning, or result analysis. To prevent it from being misinterpreted as an additional study region, we now restrict the analytical information in the text and caption to the station locations and the distinction between the development sites and the three independent test sites.

The dashed and boundary lines are likewise geographic elements of the base map. They do not represent climatic regions, dataset partitions, or model-training areas. We checked their presentation and clarified in the caption that these elements have no additional analytical meaning in this study.

**Location of revision**

*Section 2.1, Lines 137–149; Figure 1 caption, Lines 150–153.*

## Specific comment 3 Figure 2 retrieval calibration

**Reviewer comment**

Figure 2: The measurements in the panels you provide look impressive. Congratulations. Nevertheless, it would be important to know how you performed calibration (e.g. reference height) to retrieve the high-resolution backscatter coefficient (panels f–h). Please provide this information, e.g. in the respective section line 195 ff or the Annex.

**Author response**

We added descriptions of the Raman and Fernald retrieval procedures. Raman extinction does not require an assumed lidar ratio, whereas Raman backscatter is obtained from the elastic-to-Raman signal ratio with molecular normalization. When Raman retrievals are unreliable, the Fernald method is used, with a manually selected cloud-free molecular reference region typically located at 6–10 km. We also emphasize that these retrieved products are used only for auxiliary annotation and physical interpretation and are not used as model inputs.

**Location of revision**

*Section 2.3, Lines 184–217.*

**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.”*

## Specific comment 4 Section 2.2 maintenance and instrument consistency

### Reviewer comment

Sec. 2.2: It is stated that “each site performs routine maintenance and periodic calibration following operational protocols”. Can you provide more information on that or give references? Furthermore, the lidar description you give seems to be valid for all 47 sites; is this correct? Please state more explicitly.

### Author response

We clarified that the CARLNET instruments included in this study were mainly supplied by two manufacturers, but they were developed, calibrated and operated according to unified operational technical specifications. All of the 47 lidar systems are multiwavelength polarization Raman lidar with capability of 3 β +2 α +2 δ +water vapor measurements.

We also added two references documenting complementary aspects of CARLNET operation and quality assurance. Shao et al. (2025) describe the calibration and quality-control procedure for the water-vapor Raman channel, whereas Bu et al. (2026) present the network-level quality assessment of lidar raw signals, including signal saturation, Rayleigh-fit quality, effective detection range, and the identification of potential hardware anomalies.

Shao, N., Wang, Q., Bu, Z., Yin, Z., Dai, Y., Chen, Y., and Wang, X.: China Aerosol Raman Lidar Network (CARLNET)—Part I: Water Vapor Raman Channel Calibration and Quality Control, Remote Sens., 17, 414, https://doi.org/10.3390/rs17030414, 2025.

Bu, Z., Dai, Y., Mao, S., Wang, Q., Yin, Z., Wei, Y., Wang, X., Chen, Y., and Zhang, P.: China Aerosol Raman Lidar Network (CARLNET)—Part II: Quality Assessment of Lidar Raw Data, Remote Sens., 18, 663, https://doi.org/10.3390/rs18040663, 2026.

### Location of revision

*Section 2.2, Lines 154–174.*

## Specific comment 5 Missing unit after 407

### Reviewer comment

Line 158: “and 407 Raman channels” missing “nm” after “407”, please add this.

### Author response

We added the missing unit and standardized the wavelength units for all channels throughout the manuscript.

**Location of revision**

*Section 2.2, Lines 158–162: "the 407 nm water-vapor Raman channel".*

## Specific comment 6 One-month test period

**Reviewer comment**

Test data (line 236): Why was only one month chosen as the test dataset, rather than data distributed over a whole year?

**Author response**

We added the rationale for selecting a fixed continuous one-month period as the test set. The primary objective of this study is to evaluate model performance throughout continuous operational observations under natural weather evolution, day–night background variations, changes in instrument status, and different aerosol environments, rather than to select specific cloud events. A continuous one-month dataset preserves these real-world variations and avoids the event-selection bias that could arise from manually choosing representative cases.

Continuous observations from January 2025 were used at all three test sites, providing a consistent temporal basis for cross-site comparison. At the same time, we agree that a one-month period cannot represent the complete seasonal cycle. The revised manuscript therefore explicitly identifies annual and cross-season generalization as limitations of the present study and as priorities for future evaluation.

**Location of revision**

*Section 2.1, Lines 137–149; Conclusion and Discussion, Lines 700–705.*

**Relevant revised text**

*"A continuous fixed-period test set was chosen to provide an event-independent operational evaluation while preserving natural variations in cloud conditions, aerosol background, daytime–nighttime SNR, and instrument status. Using the same period at all three sites also enables a consistent cross-site comparison."*

## Specific comment 7 Radiosonde availability

### Reviewer comment

Sec. 2.5. Radiosondes available at all sites? Please clarify.

### Author response

We thank the reviewer for raising this point. Co-located operational radiosonde observations are available at all 47 CARLNET sites, but the present study does not use radiosonde data from all of them. To maintain consistency with the strict cross-site testing protocol, the radiosonde-based consistency assessment uses only observations collected at the three independent test sites during the continuous test period in January 2025. After temporal matching and quality control, 47 valid lidar–radiosonde matched cases were retained. Thus, “47” refers to the total number of matched cases aggregated across the three test sites, rather than to the number of radiosonde stations.

We have clarified the overall availability of the radiosonde observations, their actual use in this study, and the matching procedure in Sections 2.5 and 4.3. Table 4 summarizes the evaluation results for the three test sites. Given the horizontal drift of radiosonde balloons and the temporal evolution of clouds, radiosonde-derived moist layers are used only as an independent structural-consistency reference and are not treated as absolute ground truth for lidar-derived cloud boundaries.

### Location of revision

*Section 4.3, Lines 598–605; Table 4 caption, Lines 649–650.*

## Specific comment 8 Figure 3 pooling and dimensions

### Reviewer comment

Figure 3 (pipeline) – The show graph has great potential but still needs some more clarification: Also, the caption must be extended. The captions describing the pooling operation (all the way to the left) needs to be rechecked as I believe there are several typos here. Firstly, both the uppermost and the middle pooling approach are stated to be applied for “Range gate > 1600”. However, the middle approach should be applied from “800 < Range gate < 1600” from the description in the appendix. Please fix this. I would also recommend to include the actual height as an additional information instead of only the range gates, i.e., 0-12km, 12-24km, and 24-30km respectively for the different pooling

approaches (from the lowermost to the uppermost). Secondly, it is stated in the uppermost approach that a pooling of 16 bins is used. However, in the Appendix, a pooling of 8 bins is stated for the same region. Please fix this inconsistency.

**Author response**

The caption of Fig. 3 has been substantially expanded. We rechecked and standardized the vertical resampling settings across Fig. 3, Section 3.2, and Appendix A1. Each original profile contains 2048 vertical range bins at 15 m resolution. The original sampling is retained from 0 to 12 km, while resampling factors of 4 and 8 are applied at 12–24 and 24–30 km, respectively. The resulting vertical dimension is therefore 1056 rather than 512. The duplicated range labels and the incorrect factor of 16 in the original figure have also been corrected.

We now consistently refer to this operation as “vertical resampling” and clarify that it is applied before the data enter the network. Its purpose is to preserve low-level cloud boundaries and aerosol–cloud transition details while reducing redundant sampling at middle and upper levels and representing weak high-altitude structures over larger effective vertical scales. This input-preprocessing operation is distinct from the internal downsampling used by the network for feature extraction.

**Location of revision**

*Figure 3; Section 3.2, Lines 343–352; Appendix A1, Lines 719–742.*

## Specific comment 9 Adaptation ablation

**Reviewer comment**

Lines 397–398: “Any experiment that introduces a small amount of labeled data from the test sites is explicitly treated as an adaptation ablation rather than as part of the main validation protocol.” What is meant by “adaptation ablation”? I do not get this. Please consider rephrasing this point to make it clearer for the reader.

**Author response**

We removed the term “adaptation ablation” and consistently replaced it with “test-site adaptation experiment.” We also clarify that any experiment involving a small amount of labeled data from the test sites is reported separately and is not part of the main strict cross-site evaluation protocol.

**Location of revision**

*Section 4, Lines 438–447; Section 4.1, Lines 453–479.*

**Relevant revised text**

*“To evaluate the performance of different training strategies under limited manual annotations, we conducted a test-site adaptation experiment using varying amounts of labeled training data.”*

## Specific comment 10 Figure 4 introduction order

**Reviewer comment**

Lines 402–403: The x- and y-axis of Figure 4 are introduced before the actual figure is introduced. Please move this to after you introduce the figure.

**Author response**

We revised the paragraph order so that Fig. 4 and the four training strategies are introduced before the meanings of the horizontal and vertical axes are explained.

**Location of revision**

*Section 4.1, Lines 453–459.*

## Specific comment 11 No pretrain label

**Reviewer comment**

No Pretrain: The evaluation category “no pretrain” sounds a bit strange. It could be worth renaming it to “no pretraining” to be more consistent with the second category “pretraining”.

**Author response**

We agree that the original label “no pretrain” was nonstandard and have therefore removed it. We did not replace it directly with “no pretraining” because that expression indicates only the absence of pretraining without specifying the actual training strategy used for the control model. This model is initialized randomly and trained end-to-end using only expert annotations. Therefore, “direct supervised training” provides a more accurate description and forms a clear contrast with the “physics-guided pretraining” comprising Stages 1 and 2. In tables where “Pretraining” appears as a binary setting, “Yes/No” is used to indicate whether pretraining is applied.

**Location of revision**

*Figure 4 and Table 2; Section 4.1, Lines 453–479.*

## Specific comment 12 Table 2 marginal differences

**Reviewer comment**

Table 2 – Many values are only marginally different compared to others, yet the text emphasizes the benefit of, e.g., including test-site labels. Please clarify and probably rephrase to avoid overstating this effect.

**Author response**

We revised the interpretation of these results. The manuscript no longer claims that the inclusion of a small amount of test-site annotations produces a marked improvement. Instead, it states that the F1 score does not increase further after adding limited test-site annotations and that the main benefit arises from the transferable structural representations learned during pretraining.

**Location of revision**

*Section 4.2.1, Lines 482–510.*

## Specific comment 13 F1 interpretation and precision

**Reviewer comment**

F1: As you use the F1 score as the main metric, the interpretation of this value should be explained in the beginning. It would also be of interest to which extent differences are really significant. I.e., 0.928 vs. 0.937 seem to be the same for me and not a significant difference, but maybe I am wrong. Please clarify in the manuscript.

**Author response**

We added the definitions and interpretations of precision, recall, F1, and IoU in Appendix A7. Precision indicates the reliability of pixels predicted as cloud, whereas recall indicates the proportion of reference cloud pixels that are successfully detected. The F1 score is the harmonic mean of precision and recall and therefore summarizes the balance between false detections and missed cloud pixels. This interpretation is now stated explicitly when F1 first appears in the Abstract. IoU provides a stricter assessment of spatial overlap between the predicted and reference cloud masks.

Regarding the difference between 0.928 and 0.937, we agree that it cannot be described as statistically significant without repeated training experiments and uncertainty estimates. The main text therefore characterizes it as a modest deterministic improvement under a fixed dataset split and training configuration. Its operational relevance is not determined by F1 alone but is evaluated together with detectable cloud-top errors, cloud-detection rates, boundary continuity in complex cases, and structural consistency with radiosonde-derived moist layers.

**Location of revision**

*Section 4.2.1, Lines 491–500; Appendix A7, Lines 836–854.*

**Relevant revised text**

*"Precision measures the reliability of predicted cloud pixels, whereas recall measures the proportion of reference cloud pixels successfully detected. The F1 score is their harmonic mean and therefore summarizes the balance between false detections and missed clouds."*

## Specific comment 14 Interpretation of training cost

**Reviewer comment**

Lines 421–422: "Overall, the results suggest that pretraining not only reduces reliance on large-scale manual annotations for cloud-layer identification, but also provides a more controllable training cost for cross-site deployment." I cannot follow this conclusion. Can you guide more with the interpretation of your results?

**Author response**

We thank the reviewer for raising this point. In the original sentence, "more controllable training cost" was intended to refer not to computational expense, training time, or financial cost, but to the expert-annotation effort required when deploying the model at a new CARLNET site. For continuous lidar observations, unlabeled measurements and outputs from the conventional VDE algorithm are relatively easy to obtain. In contrast, pixel-level expert annotation requires the joint interpretation of multi-channel signals, polarization characteristics, and auxiliary optical products, making it time-consuming and difficult to perform at large scale. We even developed a GUI for manual annotation. But it still takes 10-60 min for annotating one-day lidar measurements, considering the many round checks of

multiple data inputs. The three-stage strategy therefore first learns time–height structural features from unlabeled observations, then introduces weak physical guidance through VDE-derived probabilistic outputs, and finally performs supervised fine-tuning using only a limited number of expert labels. Its practical significance lies in reducing dependence on large-scale manual annotation and additional site-specific labels, rather than in reducing computational cost.

The data-size experiment in Section 4.1 supports this interpretation. With only 10 labeled training days, direct supervised training achieves an F1 score of approximately 0.75, whereas the complete Stage 1 + Stage 2 pretraining strategy increases the F1 score to approximately 0.91. Stage 1 and Stage 2 individually also outperform direct supervised training. As the amount of labeled data increases, the performance of all strategies gradually approaches a plateau, indicating that the principal benefit of pretraining occurs under label-limited conditions. Furthermore, in the main strict cross-site evaluation, the model achieves an F1 score of 0.937 without using test-site observations during pretraining or test-site annotations during fine-tuning. This result further indicates that site-specific manual annotations are not a prerequisite for achieving strong cross-site performance under the current CARLNET configuration.

We agree that the original expression could be interpreted as referring to computational or financial cost, neither of which was quantified in this study. We therefore removed “controllable training cost” and limited the revised conclusion to the experimentally supported concept of “annotation efficiency”: pretraining provides its largest performance benefit under small-sample conditions and reduces the dependence of cloud-layer identification on large-scale expert annotation and additional test-site labels.

**Location of revision**

*Section 4.1, Lines 460–479.*

## Specific comment 15 Early stopping

**Reviewer comment**

Lines 426–428: You state that an early-stop strategy is used to mitigate overfitting by terminating training when validation performance does not improve for a number of consecutive epochs. Please state how many epochs were used as the threshold to stop training. This could be added in the Appendix.

**Author response**

We added the early-stopping criterion. Training was terminated when the validation F1 score failed to improve by more than ($10^{-4}$) for five consecutive epochs, and the model checkpoint with the highest validation F1 score was retained.

**Location of revision**

*Appendix A5, Lines 818–821.*

**Relevant revised text**

*"Training was terminated if the validation F1 score did not improve by more than $10^{-4}$ for 5 consecutive epochs. The model checkpoint with the highest validation F1 score was retained for subsequent evaluation."*

## Specific comment 16 Table 2 category names

**Reviewer comment**

Lines 429–430: The categories "Test sites used in pretraining" and "Test-site labels used in training" are not category names used in Table 2. Please change this to be consistent with the names used in the table.

**Author response**

We standardized the category names as "Test-site observations used in pretraining" and "Test-site annotations used in fine-tuning" and defined both categories explicitly in the main text.

**Location of revision**

*Section 4.2.1, Lines 482–490; Table 2.*

## Specific comment 17 Table 3 cloud-base bias

**Reviewer comment**

Table 3 – Base-height bias is reported as –10 m for the pretrained model, which is negligible considering the vertical resolution of the lidar.

**Author response**

We agree. After recalculation, the cloud-base bias is −12.8 m, which remains smaller than the original 15 m vertical sampling interval. We now explicitly state that this bias is negligible at the native instrumental resolution.

**Location of revision**

*Section 4.2.2, Lines 536–541; Table 3.*

## Specific comment 18 Radiosonde statistics

**Reviewer comment**

Sec. 4.3: Several results from the radiosonde evaluation are stated in this section, such as DL and VDE identification rates for multilayer and first-layer clouds, without statistics backing them up. I would recommend explaining the setup in more detail and adding statistics in the form of a table.

**Author response**

We rewrote Section 4.3 to clarify that the 47 valid comparison cases were obtained from the three independent test sites. Lidar results within ±1 h of each radiosonde launch were compared with the radiosonde-derived moist layers. We also added Table 4 to summarize the number of matched cases, structural consistency rate, and the MAEs of cloud base and detectable cloud top.

**Location of revision**

*Section 4.3, Lines 598–640; Table 4, Lines 649–650.*

## Specific comment 19 Appendix methodological details

**Reviewer comment**

Appendix:

• 611: Please specify in more detail what is meant with a “robust MAD scheme”.

• 623: “512 bins” - I believe this may be a typo. By doing the calculations myself (assuming 8 bin pooling in the uppermost approach) I end up with 1056 bins, as stated in figure 3, please check this. Additionally, it should be clearer stated that this pooling operation refers to an operation done on the input data before passing it to the model to avoid confusion with the pooling done in the actual model.

• A2: please explain in more detail the numbers you give. How have they been selected, how can they be interpreted, etc.

E.g. what does it mean: gaussian noise with probability 0.3 and SD 0.5? But also the other values are too briefly explained for full understanding and in the appendix you have enough space to do so – so no need to leave out information.

• Data augmentation – The augmentation policy (random cropping, Gaussian noise, scaling, dropout) is described, but the impact of each component on final performance is not quantified.

• 644: What is a "multi-task reconstruction objective"

• 650: What is "BCE-with-logits"? – I would guess BCE stands for Binary cross-entropy but this is never explained.

• A5: This section is hardly understandable. Please explain all the numbers you give in a proper way and also extend the description of the methodology. E.g. what is a "cosine annealing schedule"…

• 673: Are the probability thresholds referring to the soft labels? This is unclear. Please describe it more unambiguously.

**Author response**

We thank the reviewer for carefully examining the reproducibility details in the Appendix. We made the following revisions:

1. We added the formula for robust median absolute deviation (MAD) normalization and explained that the factor 1.4826 converts the MAD into a robust scale estimate consistent with the standard deviation under a normal distribution, while $(\varepsilon)$ prevents division by zero.
2. We corrected the vertical input dimension to 1056 and clarified that the resampling over 0–12, 12–24, and 24–30 km is applied before the data enter the network rather than as pooling within the network architecture.
3. We explained each data-augmentation parameter individually. The augmentation probability denotes the probability that a particular operation is applied to an individual training sample. The Gaussian-noise amplitude is interpreted relative to the channel-normalized scale, while the scaling, cropping, and dropout ranges are intended to simulate variations in signal energy, localized data loss, and channel instability. These ranges were selected empirically using observations and validation

performance from the development sites; data from the test sites were not used for parameter selection.

4. We expanded the Stage 1 “multi-task reconstruction objective” into three explicit tasks: full multi-channel reconstruction, channel-masked prediction, and temporal-frame prediction. The combined loss function and corresponding weights are now provided.
5. We clarify that BCE-with-logits is a binary segmentation loss that combines the sigmoid operation with binary cross-entropy in a single numerically stable calculation. We also provide the probabilistic supervision formulation used in Stage 2.
6. We added explanations of the cosine-annealing learning-rate schedule, optimizer parameters, and early-stopping criterion.
7. We distinguished the roles of the three thresholds. The threshold of 0.4 is used for the aerosol-suppression constraint in Stage 2; 0.5 is used only to monitor validation metrics during training; and 0.65 is the final inference threshold selected using the held-out validation subset after Stage 3 and fixed before testing. The VDE teacher probabilities in Stage 2 are always used as soft targets and are not binarized before loss calculation.

We did not add separate ablation experiments for every data-augmentation operation and therefore cannot quantify the individual contribution of each augmentation component to the final performance. The revised manuscript explicitly describes these operations as empirical robustness-enhancement settings within the complete training framework and does not attribute performance gains to any individual augmentation component.

**Location of revision**

*Appendix A1, Lines 719–742; Appendix A2, Lines 744–760; Appendix A4–A6, Lines 771–834.*

**Relevant revised text**

*“During training, a stride of 10 is used to reduce strong overlap among adjacent samples and improve computational efficiency. During inference, a stride of 1 is adopted to generate continuous cloud masks without temporal gaps.”*
*“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.”*

### Technical comment 1 Figure subpanels

**Reviewer comment**

Some figure subpanels (e.g. 7 and 8 d–e) are not referred to in the text.

**Author response**

We rechecked the panel numbering and added or standardized references to the representative subpanels in Section 4.3. The figure captions were also expanded to explain the individual panels and line styles.

**Location of revision**

*Section 4.3, Lines 598–640; Figures 7–8 captions, Lines 644–656.*

### Technical comment 2 Duplicate He et al. 2021 references

**Reviewer comment**

There are two He et al. 2021 references which significantly differ. Please separate them as 2021a and 2021b.

**Author response**

We rechecked and corrected the reference list and in-text citations. The revised manuscript retains only the He et al. (2021) reference that is actually cited and corresponds to the relevant content in the main text. Therefore, two publications from the same year no longer need to be distinguished as 2021a and 2021b.

**Location of revision**

*References and citations were checked throughout; the retained citation appears in Section 3.3, Line 386.*

### Technical comment 3 Station numbers

**Reviewer comment**

The numbers given for the three reference stations are not explained. What do these numbers stand for, e.g. 58847 (Fuzhou)?

**Author response**

We clarified that the five-digit numbers preceding the station names are traditional World Meteorological Organization (WMO) station index numbers. These internationally standardized identifiers are used to distinguish meteorological observing stations. This explanation has been added to Section 2.1 and the caption of Fig. 1. For reference, the station list used in this study is available here, while official WMO station metadata can be accessed through WMO OSCAR/Surface (Observing Systems Capability Analysis and Review Tool (OSCAR)).

**Location of revision**

*Section 2.1, Lines 137–140; Figure 1 caption, Lines 150–153.*

## Technical comment 4 Input-channel correction

**Reviewer comment**

Line 165: I think there is a mix-up. It should be the 355-parallel and the 1064-total elastic channel, correct?

**Author response**

Yes. We now provide the complete list of receiver channels and clarify in the Appendix that the model inputs comprise the 1064 nm total elastic signal, the 355 and 532 nm parallel-polarized elastic signals, and the 355 and 532 nm volume depolarization ratios.

**Location of revision**

*Section 2.2, Lines 155–170; Appendix A3, Lines 762–769.*

## Technical comment 5 Repeated transition word

**Reviewer comment**

Line 254: “Therefore, radiosonde-based …”. Consider changing “Therefore” to “Thus” to avoid using “Therefore” to start two consecutive sentences.

**Author response**

We rewrote the transition and now use “Consequently” to introduce the statement that radiosonde-derived moist layers are used only as an independent consistency reference.

**Location of revision**

*Section 2.5, Lines 270–272.*

### Technical comment 6 Deep model wording

**Reviewer comment**

Line 282: “... guides the Deep model” typo of “deep learning model”?

**Author response**

We corrected the expression to “deep learning model.”

**Location of revision**

*Section 2.6, Lines 295–303.*

### Technical comment 7 Segmentation head

**Reviewer comment**

Line 346: What is a “segmentation head”? Please explain it in a way that is more understandable for a non-AI expert.

**Author response**

We now explain in the Appendix that the segmentation head is the output module that converts the network features into a single-channel pixel-level cloud/non-cloud mask. It is also distinguished from the reconstruction head, which is activated only during Stage 1.

**Location of revision**

*Appendix A3, Lines 762–769.*

### Technical comment 8 Figure 6 contour

**Reviewer comment**

Fig. 6: Please check the label. I do not see green dashed contours.

**Author response**

We removed the incorrect description of green dashed contours. The caption of Fig. 6 now consistently states that the black solid line represents the pretrained model, the black dotted

line represents the direct supervised model, and the black boxes indicate regions of disagreement.

**Location of revision**

*Figure 6 caption, Lines 593–596.*

## Technical comment 9 Typographical error

**Reviewer comment**

Line 665: “hile” → “while”?

**Author response**

The typographical error has been corrected to “while,” and the relevant text in Appendix A5–A6 was rechecked.

**Location of revision**

*Appendix A5–A6, Lines 808–834.*

# Reviewer 2

## Overall evaluation

### Reviewer comment (in blue)

General Comments:

This manuscript presents a novel approach for cloud layer identification using a Raman lidar network system named CARLNET. The method utilizes multi-channel wavelengths, as well as an anisotropic encoder–decoder architecture, self-supervised pretraining, and traditional algorithm guidance. This approach aligns well with the scope of the AMT journal. Overall, this paper introduces a new automated method for cloud layer identification, which is acceptable for publication in AMT. However, some clarifications in the representation, as well as minor technical corrections, would enhance the quality of the manuscript, as mentioned in the specific comments and technical corrections section.

### Author response (in black)

We thank the reviewer for the positive assessment of the overall value of the manuscript and its suitability for AMT. In response to the specific comments, we clarified the role of the radiosonde-based assessment and added a systematic statistical summary. We also standardized the figures, tables, units, terminology, and technical descriptions in the Appendix.

### Location of revision

*Revisions were made primarily in Sections 2.5 and 4.3, Tables 2 and 4, Figures 5–8, and Appendix A5.*

## Specific comment 1 Skew T diagram

### Reviewer comment

In section 2.5, as “Radiosonde Observations and Cloud Identification,” the authors utilized a relative humidity thresholding method to define the boundaries of clouds by radiosondes. However, the authors also mentioned that mapping a single-column profile at a specific time is not feasible due to wind drift. Would it be possible to include a Skew-T diagram for one or a few days to validate the results? Specifically, this would help in identifying cloud bases, corresponding not only to relative humidity but also to temperatures and pressure levels.

### Author response

We thank the reviewer for this suggestion. We agree that a Skew-T diagram can intuitively display the vertical variations in temperature, dew-point temperature, and pressure and can therefore help illustrate the thermodynamic background of the moist layers identified from radiosonde observations. However, we did not add Skew-T diagrams for individual dates for the following reasons.

First, a Skew-T diagram and the relative-humidity thresholding method used in this study are based on the same radiosonde profile and therefore do not constitute independent sources of validation. A Skew-T diagram primarily identifies near-saturated layers through the proximity of temperature and dew-point temperature, while relative humidity is derived from the same set of thermodynamic observations. Consequently, adding a Skew-T diagram would change the presentation of the information but would not provide evidence independent of the existing radiosonde-based moist-layer identification. Temperature and pressure can characterize the thermodynamic environment of a moist layer, but they cannot independently demonstrate that cloud droplets detectable by lidar are present at the same altitude.

Second, a Skew-T diagram cannot resolve the spatial-sampling mismatch between the two observing systems. The lidar observes a fixed atmospheric column above the site, whereas the radiosonde may drift horizontally by approximately 30 km during ascent. Whether the radiosonde measurements are displayed as a relative-humidity profile or as a Skew-T diagram, they still represent an atmospheric profile sampled along the drifting balloon trajectory. Therefore, moist-layer boundaries derived from an individual sounding should not be treated as strict ground truth for lidar-derived cloud-base and cloud-top heights at the launch time.

Third, Skew-T diagrams for only one or two dates would provide qualitative illustrations at the case level and could be affected by the selected weather conditions and case-selection bias. Such examples would not represent the overall performance across the three independent test sites. To provide more systematic evidence and avoid reliance on selected cases, we instead evaluated 47 valid lidar–radiosonde matched cases from the three test sites. To account for cloud evolution and the spatial mismatch caused by balloon drift, lidar-derived cloud masks within ±1 h of each radiosonde launch were aggregated. We then evaluated the agreement between the two observing systems in terms of layer altitude, continuity, and multilayer structure.

We also further clarified the evidential role of the radiosonde comparison. Radiosonde observations identify moist layers using altitude-dependent relative-humidity thresholds,

whereas lidar identifies cloud layers associated with detectable scattering returns. The two methods therefore represent different physical quantities and sampling characteristics. Accordingly, the radiosonde results are used only as an independent reference for structural consistency rather than as ground truth for cloud boundaries. The statistics in Table 4 should likewise be interpreted as measures of structural consistency between the two observing systems rather than as strict cloud-identification accuracy. For these reasons, we consider the systematic analysis of all 47 matched cases more appropriate for the objectives of this study than the addition of a small number of Skew-T examples.

**Location of revision**

*Section 2.5, Lines 252–272; Section 4.3, Lines 598–640; Table 4, Lines 649–650.*

**Relevant revised text**

*"Consequently, radiosonde observations are not treated as a strict quantitative ground truth for lidar cloud-layer identification. Instead, they provide an independent atmospheric reference for evaluating the consistency of identified cloud-layer structures, including layer altitude, continuity, and multilayer occurrence."*

## Specific comment 2 Systematic radiosonde results

**Reviewer comment**

In Section 4.3, as "Independent Validation with Radiosonde Observations," it is recommended that the results be presented more clearly and systematically. In particular, it is suggested to include a table summarizing the results for improved clarity and ease of comparison.

**Author response**

We rewrote Section 4.3 and added Table 4. The table summarizes the number of matched cases, structural consistency rate, cloud-base MAE, and detectable cloud-top MAE for expert annotations, the proposed model, and VDE across the 47 matched cases.

**Location of revision**

*Section 4.3, Lines 598–640; Table 4, Lines 649–650.*

**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. The proposed model achieves a structural consistency rate of 38.3%, higher than VDE (34.0%) and expert annotation (27.7%).”*

### Technical correction 1 Figure and section abbreviations

**Reviewer comment**

The terms (e.g., Figure/Fig., Section/Sect.) should follow the Copernicus manuscript guidelines and be used consistently throughout the manuscript.

**Author response**

Following the Copernicus manuscript conventions, we standardized the in-text abbreviations as “Fig.” and “Sect.” throughout the manuscript.

**Location of revision**

*Terminology was standardized throughout the manuscript, including Sections 2–4 and all figure captions.*

### Technical correction 2 Units and mathematical notation

**Reviewer comment**

Lines 156–157 and 217, 220: Please use standard distances for units and mathematical signs, and ensure consistency throughout the manuscript.

**Author response**

We systematically checked and standardized the notation of wavelength, distance, altitude, resolution, and percentage units throughout the manuscript. Mathematical symbols and interval notation were also standardized.

**Location of revision**

*Section 2.2, Lines 155–174; Section 2.4, Lines 219–247; Appendix A1–A7, Lines 719–854.*

### Technical correction 3 Table 2 structure

**Reviewer comment**

In Table 2, it is recommended to reform the table using abbreviations for better structure.

**Author response**

We reorganized the column headings in Table 2 and separately indicated whether unlabeled test-site observations were used during pretraining and whether test-site annotations were used during fine-tuning. To avoid introducing unfamiliar abbreviations, the strategy names are retained in full, but the table structure has been made more compact and consistent with the main text.

**Location of revision**

*Section 4.2.1, Lines 482–500; Table 2.*

## Technical correction 4 Figure 5 caption

**Reviewer comment**

In Figure 5, it is recommended to include every annotation used in the figure within the caption as well.

**Author response**

The caption of Fig. 5 now identifies the method corresponding to each panel and defines GT, N, MAE, bias, RMSE, R, and the dashed 1:1 line.

**Location of revision**

*Figure 5 caption, Lines 559–564.*

## Technical correction 5 Figure 6 wording and contour

**Reviewer comment**

In Figure 6b, in terms of “noise misclassified as ...” is something missing after “as”? Please also define the green dashed contour, as it is not visible on the plot.

**Author response**

We completed the description of the misclassification shown in the figure and removed the reference to green dashed contours, which was inconsistent with the actual figure. The revised caption defines only the elements that are present: the black solid lines, black dotted lines, and black boxes.

**Location of revision**

*Figure 6 and its caption, Lines 593–596.*

### Technical correction 6 Typographical error

#### Reviewer comment

In A5, line 665: replace “hile” with “while”.

#### Author response

The typographical error has been corrected.

#### Location of revision

*Appendix A5, Lines 808–821.*

# 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."*