

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.

Major comments

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.

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.

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.

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.

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.

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.

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.

Minor comments

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.

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.

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.

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.