

REVIEW RESPONSE

Fernando K.I. Fugihara

Universidade Estadual de Campinas – UNICAMP

Campinas, SP, Brazil, 13083-896

f205067@dac.unicamp.br

1 ASSOCIATE EDITOR REVIEW

1.1 GENERAL COMMENT

Dear Dr. de Souza and Co-authors

Thank you for submitting your manuscript to RSE. I have now received two insightful reviews from field experts. Both reviewers find the topic relevant and see potential value in an entropy-based framework for evaluating cloud masks. However, they also identify several substantive issues that currently limit the robustness, interpretability, and impact of the work.

Based on the reviews and my own assessment, I have to recommend a rejection of the manuscript in its form. However, recognizing the potential of the manuscript, I encourage you to submit a revised manuscript if you can address the reviewers comments thoroughly and revised the manuscript accordingly Please note that further consideration of the manuscript is conditional on addressing the following non-negotiable technical points.

Reply to Editor General Comment: Dear Associate Editor Zhibo Zhang, we appreciate your thoughtful review and the constructive feedback from you and the reviewers. We thoroughly addressed the identified issues and revised the manuscript accordingly. We ensured that all non-negotiable technical corrections were included in our updated submission. Thank you for the opportunity to improve our work.

1.2 NON-NEGOTIABLE TECHNICAL CORRECTIONS

1.2.1 COMMENT #1 ✓

Consistency and correctness of the ΔH definition and interpretation: Your definition of ΔH must be internally consistent with the interpretation and conclusions throughout the manuscript. Please revise the text, figures, and discussion so that the sign and meaning of ΔH are correct and unambiguous, and ensure that any results or conclusions that depend on ΔH are updated accordingly.

Reply to non-negotiable technical corrections - comment #1: We updated ΔH definition and interpretation throughout the manuscript to ensure consistency and correctness. All dependent results and conclusions were revised accordingly.

1.2.2 COMMENT #2 X

Explicit and correct multiclass evaluation formulation: The manuscript evaluates multiple classes (e.g., thick cloud, thin cloud, shadow, clear land), but the reported metrics and formulas are largely presented in binary form. Please explicitly state how precision, recall, F1, IoU, and any other measures are computed in the multiclass setting (e.g., per-class with macro-averaging, micro-averaging, one-vs-rest, or reduction to a binary cloud/non-cloud problem). Report per-class results where appropriate, and ensure the text, tables, and captions make the evaluation protocol transparent and reproducible.

Reply to non-negotiable technical corrections - comment #2: We inserted a short snippet in the **Section 2.3** to explain how we computed the metrics in the multiclass setting to ensure transparency and reproducibility of the evaluation protocol. We also showed the metrics for all classes in the supplementary material. The multi-class formulation used by precision, recall, F1, IoU metrics were computed per class using a one-vs-rest (OvR) interpretation of the multiclass confusion matrix.

1.2.3 COMMENT #3 X

Reproducibility of the experimental setup: Please strengthen the Methods so an informed reader can reproduce the analysis. This includes the exact datasets/splits, preprocessing steps, and configuration choices used for each cloud mask method. For the SAM-based analysis, clearly specify the mode and settings used (e.g., fully automatic vs prompt-based, model variant, and key parameters that influence the number and scale of predicted objects).

Reply to non-negotiable technical corrections - comment #3: Thank you for this comment. We have revised and clarified the entire manuscript, with particular emphasis on the methodological section. We have strengthened the Methods by detailing datasets and splits, preprocessing steps, and configuration choices for each cloud mask method. For the SAM-based analysis, we previously reported that we used the fully automatic mode. We now specify additional settings and key parameters.

1.3 THERE ARE ALSO A FEW MINOR ISSUES:

1.3.1 MINOR ISSUE #1 X

Clarify the paper’s contributions in the Introduction: Conclude the Introduction with a concise, structured list of contributions, and ensure the novelty relative to existing cloud-mask evaluation practices is articulated clearly.

Reply to minor issue #1: We appreciate your feedback. The second half of the introduction was completely rewritten to address this recommendation and other reviewers’ suggestions.

1.4 MINOR ISSUE #2 X

Substantiate the “downstream impact” claim beyond object counts: If the goal is to quantify downstream impacts of cloud masking, please strengthen the evidence that the proposed proxy metric (e.g., ΔNO from SAM) is meaningfully tied to a downstream task objective. This could include additional validation on task-relevant outcomes (not just object counts), sensitivity analyses showing robustness to configuration choices, or a carefully framed scope statement that limits claims accordingly.

Reply to minor issue #2: We agree that using only object counts as a measure of segmentation quality is unconvincing. However, our focus is on the quality and performance of the cloud mask. Therefore, given the scope of this work, we decided not to conduct additional experiments involving downstream tasks. Thus, we clarified the scope statement to limit claims accordingly. **We added the following sentence in the last paragraph of the Introduction: “While we acknowledge that object count changes may not fully capture all aspects of downstream task performance, our focus remains on evaluating cloud mask quality and its potential implications for downstream tasks.” We also added the following sentence in the Conclusion section: “While we acknowledge that object count changes may not fully capture all aspects of downstream task performance, our focus remains on evaluating cloud mask quality and its potential implications for downstream tasks.”**

Our benchmark dataset CloudSEN12+ does not have land cover annotations, only cloud based types, because of that, we cannot use metrics based on other types of classes such as crops, forest, etc.

Fernando, pensei em criar uma nova métrica que relaciona ΔNO com ΔH , tipo um quociente de variação (derivada) para tentar capturar melhor o impacto na tarefa downstream. O que você acha?

1.5 MINOR ISSUE #3 X

Statistical support for threshold analyses and recommendations: Where the manuscript recommends thresholds or “optimal ranges,” please include uncertainty estimates and/or variability across scenes or subsets, and avoid over-generalized recommendations unless they are supported.

Please let me know if you have any questions.

Reply to minor issue #3: While clarifying the article, we noticed there was too much information. We then decided to relocate some details to the supplementary material and remove the threshold sensitivity analysis section. Therefore, we will not incorporate this correction.

2 REVIEWER #1

2.1 GENERAL COMMENT X

Thank you for submitting this interesting manuscript. The study presents a novel entropy-based framework for evaluating cloud mask performance and introduces segmentation-derived object loss as a proxy for downstream task impact. The motivation is relevant and the framework has potential value; however, several methodological and clarity-related aspects require substantial revision before the manuscript can be considered for publication.

Reply to General Comment:

2.2 MAJOR SUGGESTIONS

2.2.1 COMMENT #1 X

Clarification and Emphasis of Contributions in the Introduction The final paragraph of the Introduction mainly summarizes the structure but does not clearly state the specific contributions of the work. It is standard in Remote Sensing of Environment papers to conclude the introduction with a structured list of contributions. I recommend revising the paragraph to explicitly state the novel aspects of the work. For example, proposing an entropy-based framework to evaluate downstream segmentation effects, introducing SAM-derived object counts (ΔNO) as a task-oriented metric, and conducting large-scale comparisons across diverse models and datasets. This will help readers quickly grasp the scope and value of the paper.

Reply to Comment #1:

2.2.2 COMMENT #2 X

The entropy difference ΔH is defined as $H(\cdot) - H(\cdot)$, but the manuscript claims that ΔH greater than 0 indicates increased uncertainty. This statement is incorrect based on the definition provided. A positive ΔH should mean the entropy decreased after masking. This inconsistency should be corrected, and any dependent conclusions should be updated accordingly.

Reply to Comment #2:

2.2.3 COMMENT #3 X

The manuscript evaluates four classes (thick cloud, thin cloud, shadow, and land), but almost every formula provided are only defined for binary classification. It is not stated how these metrics (precision, recall, F1-score, and IoU) were extended to the multiclass setting. Please clarify whether the metrics were computed per class, then averaged (e.g., macro-averaging), aggregated across all pixels (micro-averaging), or reduced to a binary cloud vs non-cloud mask. This clarification is important to properly interpret the comparison results and ensure alignment with multiclass evaluation standards. Relevant intercomparison frameworks and multiclass cloud detection formulations are available in the recent cloud detection literature and can help standardize reporting and interpretation in a multiclass setting.

Reply to Comment #3:

2.2.4 COMMENT #4 X

The entropy analysis presented in the manuscript is based on single-band histogram evaluations (using the red band), which may be insufficient to distinguish classes with similar reflectance characteristics, such as cloud vs snow, cloud vs other bright surfaces, or shadow vs water or dark terrain.

Such spectral ambiguities are highly relevant in real-world scenes and may limit the robustness of entropy difference (ΔH) as an informative evaluation metric in heterogeneous or high-albedo environments. It is unclear whether the tested datasets include these challenging cases. If not, this should be acknowledged as a limitation, and the authors may consider discussing whether extending the entropy calculation to multi-band, joint, or texture-based forms might improve class separability and make ΔH more robust in such scenarios.

Reply to Comment #4:

2.2.5 COMMENT #5 X

The threshold experiments shown in Figure 6 and Tables 3-4 provide suggested optimal ranges for cloud probability threshold (CPT) and clear threshold (CT) values. However, the results are presented without any uncertainty estimates (e.g., variance, standard deviation, or confidence intervals). As threshold selection can be sensitive to dataset composition and the specific scenes tested, the absence of such statistical measures makes it difficult to judge the robustness and generalizability of the proposed thresholds. Including uncertainty estimates would strengthen the threshold analysis and ensure that the recommendations are not overinterpreted based on a limited or non-representative subset of test images.

Reply to Comment #5:

2.2.6 COMMENT #6 X

Several IoU values reported are below 0.5, particularly for thin cloud and shadow classes. IoU values below 0.5 are generally considered insufficient in operational scenarios. The manuscript should clarify whether this level of performance is expected for the compared methods or if the results are mainly for relative comparison. Some context from existing benchmark studies would help.

Reply to Comment #6:

2.2.7 COMMENT #7 X

The study includes comparisons using U-Net-based models and SEnSeI-v2, which are valid choices and widely used in cloud segmentation tasks. However, recent advancements in deep learning for remote sensing have shown that transformer-based architectures (e.g., Swin-Unet, SegFormer, UNetFormer, and MAE-based models) often outperform traditional convolutional-based models on various semantic segmentation problems, including cloud detection. These architectures benefit from long-range spatial dependency modeling and multi-scale context aggregation, often leading to improved performance in heterogeneous and complex environments. Since this manuscript aims to provide a comprehensive comparison of cloud masking methods at scale, it would be appropriate to either (a) include at least one representative transformer-based cloud segmentation model in the comparison, or (b) acknowledge this omission explicitly as a limitation. Including such a model (or addressing the absence clearly) would ensure that the evaluation framework and conclusions are aligned with the current state of the art (SOTA) in deep learning for remote sensing.

Reply to Comment #7:

2.2.8 COMMENT #8 X

The code repository is mentioned, but it is not clear stated whether the full set of cloud mask outputs, entropy difference (ΔH) maps, and segmentation object loss (ΔNO) visualizations used in the paper are to be included. Since the visual and per-image results play a significant role in your evaluation (e.g., Figure 3 and threshold analysis), I recommend adding these outputs directly to the

repository or submitting them as supplementary material. This would improve transparency and allow readers to exactly reproduce the evaluations and visual comparisons presented in the paper.

Reply to Comment #8:

2.3 MINOR SUGGESTIONS

2.3.1 COMMENT #1 - FIGURE 3 IMAGE SELECTION X

The selection method of images used in qualitative comparisons (e.g., Figure 3) is not described. Please clarify whether these were randomly selected, chosen to represent typical cases, or selected for specific traits such as edge cases.

Reply to Comment #1:

2.3.2 COMMENT #2 - GRAYSCALE VISUALIZATION OF CLASSES X

The grayscale representation used in qualitative figures does not clearly distinguish between thick and thin clouds, especially in printed or grayscale copies. Consider using color or distinct intensity patterns.

Reply to Comment #2:

2.3.3 COMMENT #3 - CLARIFY THRESHOLD TERMINOLOGY X

Some terms like cloud probability threshold and clear threshold should be more clearly defined, especially for algorithms like s2cloudless that use dual thresholds.

Reply to Comment #3:

2.3.4 COMMENT #4 - JUSTIFY SINGLE BAND ENTROPY FOCUS X

The manuscript often uses entropy from the red band as representative of all RGB channels. Please clarify why this choice is made or show that entropy across all bands behaves similarly.

Reply to Comment #4:

2.3.5 COMMENT #5 - CONSISTENT CITATION OF SHANNON X

There is inconsistency in the reference to Shannon's work. Please standardize whether you are referring to the 1948 paper or the 1949 book.

Reply to Comment #5:

2.3.6 COMMENT #6 - SHADOW THRESHOLD SENSITIVITY X

The threshold of NIR less than 0.1 is used to detect shadows. Please provide justification for this threshold or show sensitivity of metrics to changes in this value.

Reply to Comment #6:

2.3.7 COMMENT #7 - NOTATION CONSISTENCY X

Ensure consistent formatting for technical terms throughout the paper, such as ΔH , IoU, and class labels. Check thoroughly.

Reply to Comment #7:

2.3.8 COMMENT #8 - HIGHLIGHTING BEST RESULTS IN TABLES X

Best values in tables are currently indicated with bold or underline, but not always clearly. Consider using a clearer method or note if differences are not statistically significant.

Reply to Comment #8:

2.3.9 COMMENT #9 - FIGURE CAPTIONS COULD BE MORE INFORMATIVE X

Some figure captions do not provide enough context about what the figure is demonstrating. Adding such detail will make the visual results easier to interpret.

Reply to Comment #9:

2.3.10 COMMENT #10 - LANGUAGE AND FORMATTING X

A final proofreading pass is recommended to ensure consistency in terminology, grammar, and figure/table formatting.

Reply to Comment #10:

3 REVIEWER #2

3.1 GENERAL COMMENT X

This study attempts to integrate information entropy with segmentation models, proposing a novel approach for evaluating cloud masks. However, there are still some points that need to be clearly specified and addressed before acceptance.

Reply to General Comment:

3.2 MAJOR SUGGESTIONS

3.2.1 COMMENT #1 ✓

Overall, the contribution and innovativeness of the manuscript are insufficient.

Reply to Comment #1: We truly appreciate for your comment, although we do believe that the main idea of the paper is innovative. However, your comment showed that the innovation and contribution were not clearly explained. Therefore, we clarified these ideas in the last paragraph of the Introduction.

”””

3.2.2 COMMENT #2 X

The Abstract and Introduction sections fail to highlight the contributions, innovations, and the significant importance of the research. Furthermore, the logic is confusing in several places, with paragraphs lacking transitions and logical flow. Excessive detail is given to non-essential background information, while the description of the present work is comparatively scarce.

Reply to Comment #2:

3.2.3 COMMENT #3 X

The content in the fourth paragraph is highly repetitive and disorganized, redundantly using phrases like "increasing research," "have been proposed," and "have proposed."

Reply to Comment #3:

3.2.4 COMMENT #4 ✓

The experimental section lacks necessary explanations and details, for instance, whether prompts were used for the Segment Anything Model (SAM). Furthermore, using only the change in the number of objects as a measure of segmentation quality is unconvincing.

Reply to Comment #4: The explanation about whether we used prompts for SAM was already explained in the first paragraph of the Section 3.3 as follows

"For our analysis, we utilized the Vision Transformer-Large (ViT-L) image encoder with fully automatic segmentation mode, which segments the entire image. This differs from Wang et al. (2024) work, which utilized the Visual Reference Prompt tool to indicate the object to be segmented."

3.2.5 COMMENT #5 X

The manuscript does not discuss the practical impact of cloud masking on subsequent classification tasks, remaining solely at the level of "object count." If the aim is to discuss the impact of cloud masking on classification tasks, it is recommended to introduce and evaluate using a dataset specifically designed for classification tasks.

Reply to Comment #5:

3.2.6 COMMENT #6 X

The experiments lack validation across different geographical regions and sensors (generalization experiments), thus failing to demonstrate that the method is universally applicable.

Reply to Comment #6:

REFERENCES

Y. Wang, Y. Zhao, and L. Petzold. An empirical study on the robustness of the segment anything model (SAM). *Pattern Recognition*, 155:110685, 2024. doi: 10.1016/j.patcog.2024.110685.