REVIEW RESPONSE

Fernando K.I. Fugihara
Universidade Estadual de Campinas – UNICAMP
Campinas, SP, Brazil, 13083-896
f205067@dac.unicamp.br

1 REVIEWER #1

1.1 GENERAL COMMENT

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:

1.2 MAJOR SUGGESTIONS

1.2.1 COMMENT #1

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:

1.2.2 COMMENT #2

The entropy difference ΔH is defined as $() - ()$, 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:

1.2.3 COMMENT #3

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 multi-class setting.

Reply to Comment #3:

1.2.4 COMMENT #4

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:

1.2.5 COMMENT #5

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:

1.2.6 COMMENT #6

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:

1.2.7 COMMENT #7

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:

1.2.8 COMMENT #8

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:

1.3 MINOR SUGGESTIONS

1.3.1 COMMENT #1 - FIGURE 3 IMAGE SELECTION

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:

1.3.2 COMMENT #2 - GRAYSCALE VISUALIZATION OF CLASSES

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:

1.3.3 COMMENT #3 - CLARIFY THRESHOLD TERMINOLOGY

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:

1.3.4 COMMENT #4 - JUSTIFY SINGLE BAND ENTROPY FOCUS

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:

1.3.5 COMMENT #5 - CONSISTENT CITATION OF SHANNON

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:

1.3.6 COMMENT #6 - SHADOW THRESHOLD SENSITIVITY

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:

1.3.7 COMMENT #7 - NOTATION CONSISTENCY

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

Reply to Comment #7:

1.3.8 COMMENT #8 - HIGHLIGHTING BEST RESULTS IN TABLES

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:

1.3.9 COMMENT #9 - FIGURE CAPTIONS COULD BE MORE INFORMATIVE

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:

1.3.10 COMMENT #10 - LANGUAGE AND FORMATTING

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

Reply to Comment #10:

2 REVIEWER #2

2.1 GENERAL COMMENT

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:

2.2 MAJOR SUGGESTIONS

2.2.1 COMMENT #1

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

Reply to Comment #1:

2.2.2 COMMENT #2

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:

2.2.3 COMMENT #3

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:

2.2.4 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:

2.2.5 COMMENT #5

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:

2.2.6 COMMENT #6

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