ASSESSING THE IMPACT OF THRESHOLD ON CLOUD MASKING ALGORITHMS FOR DOWNSTREAM TASKS

Fernando K. I. Fugihara[1], Marlon F. de Souza[2], Rubens A. C. Lamparelli[1], Helio Pedrini[3]
[1]Plasticulture Engineering Center, Universidade Estadual de Campinas
[2] Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo
[3]Institute of Computing, Universidade Estadual de Campinas
f205067@dac.unicamp.br

ABSTRACT

This study presents an entropy-based threshold sensitivity analysis of cloud masking (CM) algorithms, quantifying how different threshold values affect scene uncertainty and downstream segmentation performance. We evaluate two widely used CMs, Cloud Score+ and s2cloudless, across a wide range of thresholds using the CloudSEN12+ benchmark dataset. In addition to traditional accuracy metrics, including Overall Accuracy (OA), F1-score (F1), and Intersection over Union (IoU), we considered the mean relative difference in object counts ($\overline{\Delta NO}$), derived from Segment Anything Model (SAM) outputs, and the mean entropy difference ($\overline{\Delta H}$). Our results reveal a clear trade-off between uncertainty reduction and information loss as masking becomes more aggressive, and show that threshold values that maximize pixel-level accuracy do not necessarily optimize segmentation outcomes. The primary contribution is to show that choosing the right threshold is crucial for balancing the elimination of cloud-related uncertainty with the preservation of valuable information that improves the quality of downstream tasks.

1 INTRODUCTION

Despite considerable advancements in satellite sensor quality and data availability over recent decades, cloud cover remains a significant challenge in optical remote sensing (RS). Thick clouds completely block surface reflectance, while thin clouds and shadows partially distort it, creating inconsistent observations that hinder applications requiring dense time series, such as agricultural monitoring and disaster response (Wen et al., 2001; Robinson et al., 2019; Meraner et al., 2020; Tsardanidis et al., 2025). Long-term records highlight the magnitude of the problem. King et al. (2013) reported cloud cover over 67% of the surface across 12 years of Moderate Resolution Imaging Spectroradiometer (MODIS) data. Unlike typical noise, clouds represent a structured and scene-dependent perturbation that complicates the extraction of reliable surface information (Wang et al., 2024).

A wide range of cloud masking (CM) techniques has been developed, ranging from physically based on advanced deep-learning models (Zhu & Woodcock, 2012; Foga et al., 2017; Zhu & Helmer, 2018; Pasquarella et al., 2023; Francis, 2024; Aybar et al., 2024). However, the optimal use of CM remains challenging because thin clouds, although visually degrading, often preserve enough information for modern RS foundation models to correctly identify land-cover objects (Wang et al., 2024). This highlights the critical role of threshold selection, which directly controls the aggressiveness of cloud masking. An overly aggressive mask may hide objects that could otherwise be accurately segmented, while too conservative masking retains cloud-induced noise that degrades downstream analysis. Therefore, it is crucial to balance filtering contaminated areas with preserving as much valuable data as possible.

The optimal threshold depends on the dataset and metric being analyzed. Pasquarella et al. (2023) reported that different datasets require distinct threshold values for CloudScore+ (CS+) and s2cloudless. In the Tarrio dataset (Tarrio et al., 2020), the recommended thresholds were 0.64 for CS+ and 0.12 for the s2cloudless Cloud Probability Threshold (CPT), while for the CMIX-I PixBox

dataset (Skakun et al., 2022), tested thresholds were 0.50 for CS+ and 0.40 for s2cloudless CPT. However, these analyses relied primarily on traditional confusion-matrix metrics such as overall accuracy, precision, and F1-score (Pasquarella et al., 2023), which may not fully reflect the effects of threshold selection on downstream segmentation or classification tasks.

Two metrics proposed by Fugihara et al. (2025) were used: the mean entropy difference ($\overline{\Delta H}$), which captures changes in spectral information, and the relative number of object counts ($\overline{\Delta NO}$), which reflects the effect of cloud masking on subsequent segmentation outcomes. Our main contribution is to show the impact of different cloud-masking thresholds on scene preservation, identifying the thresholds that maximize scene meaningful information. These principles can also be translated to other downstream tasks beyond full-scene segmentation.

2 Materials and Methods

2.1 Data

In this study, we used the CloudSEN12+ benchmark dataset (Aybar et al., 2024). It includes over 50,000 Harmonized Sentinel-2 (S2) Multispectral Instrument (MSI) Level-2A surface reflectance (SR) images from various parts of the world, each with a human-annotated cloud label. Access was via the Python library tacoreader v0.5.6 (Aybar, 2025). Aybar et al. (2024) classified images into four land cover (LC) categories: clear land, thick clouds, thin clouds, and cloud shadows. These are grouped into five cloud cover conditions: cloudless (0%), mostly clear (0–25%), slightly cloudy (25–65%), moderately cloudy (45–65%), and cloudy (>65%). For this analysis, all classes and conditions were included.

2.2 Cloud Mask Algorithms and Cloud Threshold Sensitivity Analysis

We used CS+ (Pasquarella et al., 2023) and s2cloudless (Zupanc, 2017) algorithms, both with adjustable thresholds for cloud masking sensitivity (See models details in Appendix A.1). We tested metrics at thresholds from 0.1 to 0.9 in steps of 0.1 to see how threshold sensitivity impacts segmentation. An aggressive mask classifies more pixels as clouds, while conservative masks may label cloudy pixels as 'clear land'.

CS+ has two assessment bands, cs and cs_cdf, which evaluate pixel usability based on surface visibility from 0 (occluded) to 1 (clear) (Pasquarella, 2023). We used cs, derived from spectral distance to a clear reference, which is more sensitive to haze and clouds. CS+ masks various atmospheric occlusions with threshold recommendations from 0.50 to 0.65 (Pasquarella, 2023).

The s2cloudless model uses a CPT, expressed as a percentage. If a pixel's cloud probability exceeds the CPT, it is classified as a cloud. Lower CPT values produce a more aggressive mask, whereas higher CPT values yield a more conservative mask. In Section 3, to compare s2cloudless results with CS+, we used the complement of CPT (Equation 1), which aligns with the CS+ clear threshold range.

$$\text{CPT}^c = 1 - \text{CPT}, \tag{1}$$

where CPT^c is the complement of CPT.

2.3 Evaluation Metrics

This study used metrics derived from the confusion matrix, including overall accuracy (OA), F1-score (F1), and Intersection over Union (IoU). It also used two metrics proposed by Fugihara et al. (2025), $\overline{\Delta H}$ and $\overline{\Delta NO}$ using cloud masking and segmentation results from Segmentation Anything Model (See model specifications in the Appendix A.2) between images with and without masks (See evaluation metrics details in Appendix A.3). CM algorithms were evaluated on 2200 images from the CloudSEN12+ dataset that were filtered within the range of 10% and 90% of thick cloud pixels, varying the cloud thresholds from 0.1 to 0.9, focusing on three classes: clear land, thick clouds, and cloud shadows. Accuracy was assessed by comparing classified pixels with manual labels. The clear land class included Earth's surface, both dry land and water.

3 Results and Discussion

The threshold values for CS+ and s2cloudless determine the tolerance of the CD, meaning how strictly the algorithm classifies pixels as clouds. Figure 1 shows the $\overline{\Delta H}$ (for R, G, B bands) and $\overline{\Delta NO}$ behavior throughout the entire threshold variation range (0.1-0.9). The tradeoff between $\overline{\Delta H}$ and $\overline{\Delta NO}$ is evident. Appropriate threshold preserves relevant image information, thereby increasing image predictability and improving segmentation quality. Conversely, neglecting a proper definition of this value can lead to undesirable results. When masking intensity is decreased, unwanted noise can interfere with subsequent segmentation. Conversely, aggressive cloud masking can remove pixels containing low noise and relevant information that a robust model could otherwise explore to improve output quality. Although we presented both tradeoff graphs with similar threshold ranges on the horizontal axis of Figure 1 to facilitate comparison, s2cloudless and CS+ adopt different thresholding mechanisms (Braaten, 2020a).

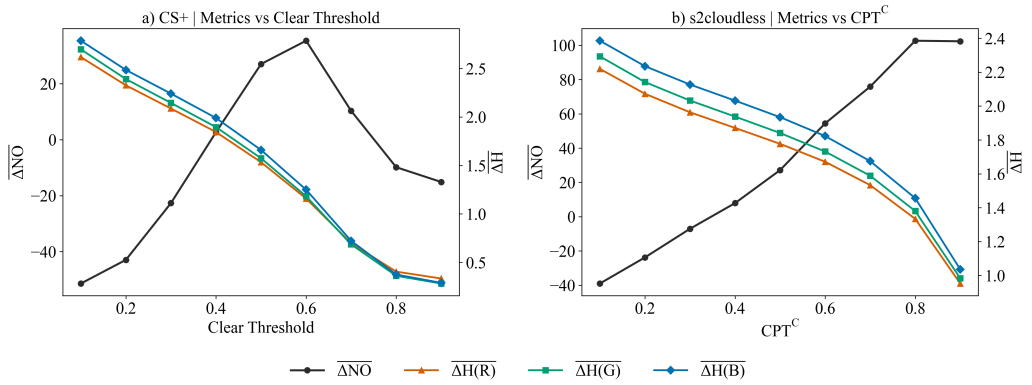

Figure 1: Trade-off between $\overline{\Delta \mathrm{NO}}$ and $\overline{\Delta \mathrm{H}}$ (for R, G, B bands) computed across threshold values. (a) CS+ results as a function of the clear threshold. (b) s2cloudless results as a function of CPT^{C}.

CS+ provides a continuous quality assessment (QA) score that can be thresholded to mask any atmospheric occlusions, such as clouds and cloud shadows (Pasquarella, 2023). Figure 1a illustrates the variation in CS+ metrics, where the relative number of object counts increased as the mean entropy difference decreased, until it reached the clear threshold of 0.6, and then the $\overline{\Delta NO}$ reduced. Thresholds from 0 to 0.3 preserved much of the original entropy and either increased somewhat or maintained the same number of objects as the raw. Conversely, thresholds between 0.7 and 1 resulted in a significant reduction in entropy and fewer segmented objects compared to the raw image. The most effective threshold ranged between 0.4 and 0.6, where, compared to raw, entropy was relatively lower, and more objects were segmented.

The trade-off between $\overline{\Delta \mathrm{H}}$ and $\overline{\Delta \mathrm{NO}}$ is similar for s2cloudless, however the decrease in the number of segmented objects between 0.7 and 1.0 does not occur (Figure 1b). We observed a stabilization at the CPT^{c} level of 0.8 (0.2 CPT). Lower complement of CPT led to more selective CD, whereas higher CPT^{c} led to more aggressive masking, where CM removed slightly cloudy scenes or unsure pixels (Braaten, 2020b). Successive entropy reduction resulted in higher $\overline{\Delta H}$ values until it reached an optimal threshold (0.10 CPT). The optimal threshold is the value that minimizes image entropy while retaining the majority of its meaningful information, thereby enabling the segmentation model to identify a greater number of objects.

In our results, the threshold values that maximize OA and F1 also differed from the thresholds that maximize $\overline{\Delta NO}$ (see Tables 1 and 2). The s2cloudless CM achieved the highest OA (0.683) and F1 at the CPT of 0.4, while the maximum $\overline{\Delta NO}$ was at 0.20 (CPT). Lowering the CTP to 0.2 (0.8 CPT^{c}) resulted in a significantly higher $\overline{\Delta NO}$ of 102.672% and a 12x increase in the number of objects, despite the reduced accuracy (OA $= 0.653$). Similarly, CS+ have different clear thresholds values that maximize OA and $\overline{\Delta NO}$. CS+ reached the highest OA (0.729) at a clear threshold of 0.5 and maximized the relative number of object counts ($\overline{\Delta NO} = 27.036\%$) in the clear threshold of 0.6, which had a lower OA (0.683).

Although OA remains an important metric, it overlooks critical aspects of satellite image analysis, such as landscape variability, making visual inspection essential despite its subjective nature. To complement visual inspection, we used metrics such as $\overline{\Delta NO}$ for segmentation consistency and $\overline{\Delta H}$ for spectral entropy analysis, providing a more comprehensive evaluation of cloud masking performance. The optimal threshold is inherently metric-dependent; therefore, a potential strategy is to select the threshold that maximizes the specific evaluation metric of interest.

Table 1: Metrics for s2cloudless evaluated across different CPT^C values, including OA, class-wise F1-scores for land, thick cloud, and shadow, $\overline{\Delta H}$ for RGB bands, $\overline{\Delta \text{NO}}$, and IoU for thick clouds (IoU results for the other classes are presented in Table A.3). Rows highlighted in gray and blue indicate the highest OA and the highest $\overline{\Delta NO}$, respectively.

CPT^c	**OA**	$\mathbf{F1_{Land}}$	$\mathbf{F1_{Thick}}$	$\mathbf{F1_{Shadow}}$	$\overline{\Delta H(R)}$	$\overline{\Delta H(G)}$	$\overline{\Delta H(B)}$	$\overline{\Delta NO}$	$\mathbf{IoU_{Thick}}$
0.10	0.606	0.633	0.700	0.131	2.327	2.418	2.526	-39.027	0.539
0.20	0.660	0.699	0.753	0.265	2.180	2.267	2.376	-23.794	0.604
0.30	0.678	0.711	0.776	0.318	2.071	2.158	2.267	-7.076	0.634
0.40	0.683	0.714	0.786	0.333	1.978	2.063	2.172	7.890	0.647
0.50	0.681	0.712	0.787	0.337	1.884	1.966	2.075	27.158	0.649
0.60	0.676	0.707	0.783	0.338	1.778	1.856	1.963	54.437	0.643
0.70	0.668	0.699	0.773	0.338	1.640	1.713	1.815	75.892	0.630
0.80	0.653	0.687	0.753	0.337	1.441	1.505	1.596	102.672	0.604
0.90	0.625	0.665	0.710	0.337	1.059	1.107	1.176	102.354	0.551

Table 2: Metrics for CS+ evaluated across different clear threshold values, including OA, class-wise F1-scores for land, thick cloud, and shadow, $\overline{\Delta H}$ for RGB bands, $\overline{\Delta \text{NO}}$, and IoU for thick clouds (IoU results for the other classes are presented in Table A.4). Rows highlighted in gray and blue indicate the highest OA and the highest $\overline{\Delta NO}$, respectively.

Clear Threshold	**OA**	$\mathbf{F1_{Land}}$	$\mathbf{F1_{Thick}}$	$\mathbf{F1_{Shadow}}$	$\overline{\Delta H(R)}$	$\overline{\Delta H(G)}$	$\overline{\Delta H(B)}$	$\overline{\Delta NO}$	$\mathbf{IoU_{Thick}}$
0.10	0.522	0.599	0.569	0.001	2.725	2.823	2.927	-51.400	0.398
0.20	0.606	0.649	0.738	0.011	2.434	2.516	2.626	-42.941	0.585
0.30	0.666	0.702	0.809	0.079	2.195	2.270	2.381	-22.664	0.679
0.40	0.711	0.761	0.824	0.197	1.952	2.021	2.130	2.670	0.701
0.50	0.729	0.811	0.799	0.295	1.641	1.699	1.800	27.036	0.666
0.60	0.683	0.762	0.743	0.330	1.266	1.307	1.390	35.374	0.591
0.70	0.562	0.491	0.662	0.335	0.806	0.811	0.862	10.317	0.495
0.80	0.470	0.157	0.616	0.332	0.512	0.486	0.515	-9.831	0.445
0.90	0.447	0.047	0.605	0.331	0.441	0.407	0.429	-15.109	0.434

4 CONCLUSION

This study conducted a systematic examination of how CM thresholds influence the effectiveness of cloud masking and the quality of downstream satellite image segmentation. Our main contribution is to demonstrate that threshold selection is critical for balancing the removal of cloud-induced uncertainty with the preservation of useful surface information.

For both CS+ and s2cloudless, a clear trade-off emerged: as masking becomes more aggressive, image entropy decreases, indicating reduced spectral uncertainty, but overly strict thresholds also remove valid surface details, degrading segmentation performance. Conversely, conservative thresholds retain more information but allow residual cloud noise to impair model predictions.

Although full scene segmentation was used as an auxiliary model to evaluate downstream effects, the underlying principle generalizes to other vision tasks, such as semantic segmentation and change detection, which can be explored in future works. In these tasks, reducing scene uncertainty while preserving meaningful structure is critical to achieving reliable, stable model performance across diverse remote sensing applications. The scope of this study was limited to a single downstream task. Evaluation was restricted to RGB imagery to maintain compatibility with SAM's input requirements.

REFERENCES

C. Aybar. Tacoreader, version 2.4.16, 2025. URL https://pypi.org/project/tacoreader/. 2

C. Aybar, L. Bautista, D. Montero, J. Contreras, D. Ayala, F. Prudencio, J. Loja, L. Ysuhuaylas, F. Herrera, K. Gonzales, J. Valladares, L. A. Flores, E. Mamani, M. Quiñonez, R. Fajardo, W. Espinoza, A. Limas, R. Yali, A. Alcántara, and L. Gómez-Chova. CloudSEN12+: The largest dataset of expert-labeled pixels for cloud and cloud shadow detection in Sentinel-2. *Data in Brief*, 56: 110852, 2024. doi: 10.1016/j.dib.2024.110852. 1, 2, 7

J. Braaten. Sentinel-2 cloud masking with s2cloudless. Google Earth Engine Tutorials, 2020a. Available at https://developers.google.com/earth-engine/tutorials/community/sentinel-2-s2cloudless. 3

J. S. K. I. S. Braaten. More accurate and flexible cloud masking for Sentinel-2 images. Medium, November 2020b. Available at https://medium.com/google-earth/more-accurate-and-flexible-cloud-masking-for-sentinel-2-images-766897a9ba5f. 3

S. Foga, P. L. Scaramuzza, S. Guo, Z. Zhu, R. D. Dilley, T. Beckmann, G. L. Schmidt, J. L. Dwyer, M. Joseph Hughes, and B. Laue. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sensing of Environment*, 194, 2017. doi: 10.1016/j.rse.2017.03.026. 1

A. Francis. Sensor independent cloud and shadow masking with partial labels and multimodal inputs. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024. doi: 10.1109/TGRS.2024.3391625. 1

F. Fugihara, M. de Souza, R. Lamparelli, and H. Pedrini. An entropy-based assessment of cloud mask algorithms for satellite image segmentation. SSRN, October 2025. Available at https://ssrn.com/abstract=5606596. 2

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.00371. 7

A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166, 2020. doi: 10.1016/j.isprsjprs.2020.05.013. 1

V. J. Pasquarella. All clear with cloud score+. Medium, October 2023. Available at https://medium.com/google-earth/all-clear-with-cloud-score-bd6ee2e2235e. 2, 3

V. J. Pasquarella, C. F. Brown, W. Czerwinski, and W. J. Rucklidge. Comprehensive quality assessment of optical satellite imagery using weakly supervised video learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2023–June, 2023. doi: 10.1109/CVPRW59228.2023.00206. 1, 2, 7

T. R. Robinson, N. Rosser, and R. J. Walters. The spatial and temporal influence of cloud cover on satellite-based emergency mapping of earthquake disasters. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-49008-0. 1

S. Skakun, J. Wevers, C. Brockmann, G. Doxani, M. Aleksandrov, M. Batič, D. Frantz, F. Gascon, L. Gómez-Chova, O. Hagolle, D. López-Puigdollers, J. Louis, M. Lubej, G. Mateo-García, J. Osman, D. Peressutti, B. Pflug, J. Puc, R. Richter, and L. Žust. Cloud mask intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 274:112990, 2022. doi: 10.1016/j.rse.2022.112990. 2, 7

K. Tarrio, X. Tang, J. G. Masek, M. Claverie, J. Ju, S. Qiu, Z. Zhu, and C. E. Woodcock. Comparison of cloud detection algorithms for Sentinel-2 imagery. *Science of Remote Sensing*, 2, 2020. doi: 10.1016/j.srs.2020.100010. 1

I. Tsardanidis, A. Koukos, V. Sitokonstantinou, T. Drivas, and C. Kontoes. Cloud gap-filling with deep learning for improved grassland monitoring. *Computers and Electronics in Agriculture*, 230: 109732, 2025. doi: 10.1016/j.compag.2024.109732. 1

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. 1

G. Wen, R. F. Cahalan, S. C. Tsay, and L. Oreopoulos. Impact of cumulus cloud spacing on Landsat atmospheric correction and aerosol retrieval. *Journal of Geophysical Research Atmospheres*, 106 (D11), 2001. doi: 10.1029/2001JD900159. 1

X. Zhu and E. H. Helmer. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. *Remote Sensing of Environment*, 214, 2018. doi: 10.1016/j.rse.2018.05.024. 1

Z. Zhu and C. E. Woodcock. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment*, 118, 2012. doi: 10.1016/j.rse.2011.10.028. 1

A. Zupanc. Improving cloud detection with machine learning. Medium, December 2017. URL `https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13`. Available at Medium.Com. 2, 7

ACKNOWLEDGMENTS

Braskem S.A. and the São Paulo Research Foundation (FAPESP) financed this work within the Plasticulture Engineering Center (grant numbers: 2021/05251-8; 2024/06854-6; 2024/06856-9). Author R.A.C. Lamparelli received financial support from the Brazilian National Council for Scientific and Technological Development – CNPq (grant number 305412/2023-0).

CODE AND DATA AVAILABILITY

The implementation code and datasets used in this study are publicly available in the following repository: `https://github.com/fekenzofugi/Threshold-Cloud-Analysis`.

A Appendix

A.1 Cloud Mask Algorithm Specifications

A.1.1 Cloud Score+

Cloud Score+ (CS+) is a single-scene ML-based CM that computes pixel-level cloud probability scores (0-100) for Landsat and S2 imagery (Pasquarella et al., 2023). The method employs clear thresholding to distinguish cloudy pixels, with CS+ incorporating temporal deep learning to enhance detection accuracy. This model does not detect thin clouds and shadows.

A.1.2 s2cloudless

S2cloudless is a single-scene ML-based CM for S2 (Zupanc, 2017). The s2cloudless was trained on a large, globally covering dataset. It is monotemporal, does not consider any spatial context, and can therefore be run at any resolution (Skakun et al., 2022). Users can convert the cloud probability map into a CM by thresholding the cloud probability map. This model does not detect thin clouds and shadows.

A.1.3 Visual Inspection of Cloud Masks

In Figure A.2, we present a Sentinel-2 scene from the CloudSEN12+ dataset (Aybar et al., 2024) along with the manually labeled (Aybar et al., 2024), CloudScore+ (Pasquarella et al., 2023) and s2cloudless (Zupanc, 2017) cloud masks for comparison.

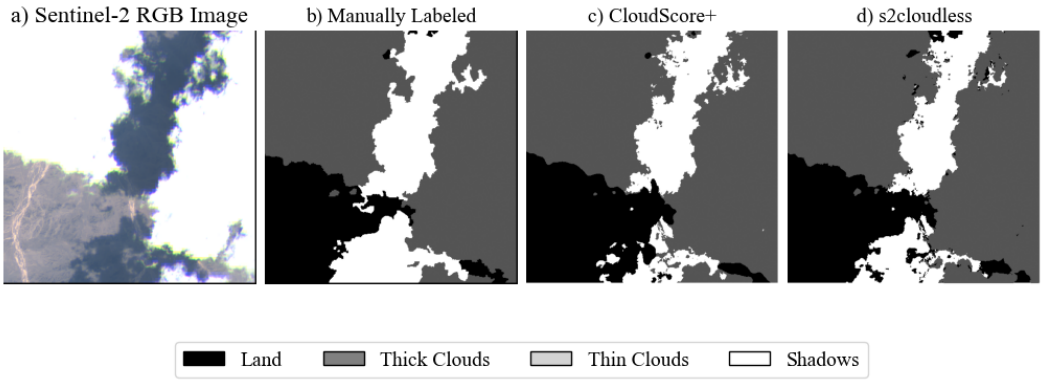

Figure A.2: Visual comparison of cloud masks for the same scene. (a) Sentinel-2 RGB image, (b) manually labeled cloud mask from the CloudSEN12+ dataset, (c) CloudScore+ cloud mask, (d) s2Cloudless cloud mask. The comparison highlights differences between automated methods and human-labeled reference data.

A.2 Segmentation Model Specifications

This study used SAM (Kirillov et al., 2023) for automatic image segmentation. SAM, trained on the SA-1B dataset of over 1 billion masks across 11 million images, performs zero-shot segmentation without task-specific training. We used the Vision Transformer-Large (ViT-L) encoder in fully automatic mode to segment the entire image. Since SAM operates exclusively on RGB inputs, our analysis was restricted to the RGB bands. Consequently, entropy computation for additional spectral bands was not considered, as they are not utilized by the segmentation model.

We used SAM as an aid to evaluate whether CM enhances the self-information quality of an image, making it more predictable. The results from SAM were assessed by visual inspection and by counting segmented objects in aerial images with different levels of cloud cover.

A.3 Evaluation Metrics

Various methodological approaches were chosen to enhance the comparison. For the CM comparison, metrics based on the confusion matrix were used, such as Overall Accuracy (OA), F1-score, and Intersection over Union (IoU). Additionally, we introduced two new metrics ($\overline{\Delta H}$ and $\overline{\Delta NO}$), which assess the relative change in segmentation results between images with masks and those without. The evaluation of CM algorithms was performed using 2200 images from the CloudSEN12+ dataset that were filtered within the range of 10% and 90% of thick cloud pixels. We focused on four classes: clear land, thick clouds, thin clouds, and cloud shadows from CloudSEN12+ and measured accuracy by comparing each classified pixel to its manually labeled reference. The "clear land" class encompasses Earth's surface, including dry land and water.

OA formulation is presented in Equation A.2.

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{A.2}$$

where TP denotes the quantity of true positives, TN represents the count of true negatives, FP refers to the number of false positives, and FN indicates the quantity of false negatives.

IoU and F1-Score were calculated for each class using a one-vs-rest (OvR) approach based on the multiclass confusion matrix. IoU formulation is presented in Equation A.3, and F1-Score formulation is presented in Equation A.4).

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{A.3}$$

$$\text{F1}_{score} = \frac{2}{\text{UA}^{-1} + \text{PA}^{-1}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{A.4}$$

where UA stands for User's Accuracy and PA stands for Producer's Accuracy.

The relative difference in object counts (ΔNO) mathematical formulation is presented in Equation A.5.

$$\Delta NO = \frac{\text{Nº OBJ}_M - \text{Nº OBJ}_R}{\text{Nº OBJ}_R} \times 100 \tag{A.5}$$

where Nº OBJ_M represents the total number of objects segmented in the masked image, and Nº OBJ_R indicates the total number of objects segmented in the corresponding raw image.

Equation 9 shows the entropy difference between the raw and masked images.

$$\Delta H(Band) = H(Band)_{Raw} - H(Band)_{Masked} \tag{A.6}$$

In the analysis, we used the mean values for ΔNO and ΔH across the dataset, which we denoted as $\overline{\Delta NO}$ and $\overline{\Delta H}$.

A.4 Additional Results

Tables A.3 and A.4 present supplementary results that complement the metrics shown in Tables 1 and 2, respectively.

Table A.3: IoU for s2cloudless evaluated across different CPT^C values for land, thick cloud and shadow classes.

$\textbf{CPT}^C$	$\textbf{IoU}_{\textbf{Land}}$	$\textbf{IoU}_{\textbf{Thick}}$	$\textbf{IoU}_{\textbf{Shadow}}$
0.10	0.463	0.539	0.070
0.20	0.537	0.604	0.152
0.30	0.552	0.634	0.189
0.40	0.555	0.647	0.200
0.50	0.552	0.649	0.202
0.60	0.547	0.643	0.203
0.70	0.538	0.630	0.203
0.80	0.523	0.604	0.202
0.90	0.499	0.551	0.202

Table A.4: IoU for CS+ evaluated across different clear threshold values for land, thick cloud and shadow classes.

Clear Threshold	$\textbf{IoU}_{\textbf{Land}}$	$\textbf{IoU}_{\textbf{Thick}}$	$\textbf{IoU}_{\textbf{Shadow}}$
0.10	0.427	0.398	0.001
0.20	0.481	0.585	0.005
0.30	0.540	0.679	0.041
0.40	0.614	0.701	0.109
0.50	0.682	0.666	0.173
0.60	0.615	0.591	0.197
0.70	0.325	0.495	0.201
0.80	0.085	0.445	0.199
0.90	0.024	0.434	0.198