This paper applies a RF classifier to ground-based sky images from several locations for estimating cloud fraction. The authors prepare annotated datasets, train site-specific and merged RF models, and compare the model's CF output against TSI results. While the manuscript is well-organised and the topic is of interest for the atmospheric observation and machine learning communities, I find the work lack of methodological novelty, analysis depth, and evaluation rigor. The comments below aim to guide the authors toward a significantly strengthened version of this manuscript.

We sincerely appreciate the reviewer's critical assessment and encouragement towards the technical rigor of our study. We also understand the concerns raised about the evaluation of the model and better quantification of the biases. Accordingly, we have addressed all the concerns and our response to the specific comments, documented below. The **bold text in black represents the reviewer's comments**, followed by our responses in blue colour regular text, and the *italic texts in red indicate the corresponding changes made in the main manuscript*.

---

1. The model is explicitly trained at the pixel level (as shown in Fig. 3), yet the evaluation is based solely on cloud fraction (CF), a scene-level aggregate statistic. This disconnect is concerning. If the model is trained to perform per-pixel segmentation, why are there no pixel-wise metrics (e.g., accuracy, F1, precision/recall, IoU) reported? This omission makes it difficult to assess how well the model actually distinguishes cloud vs. sky on a per-pixel basis, and not just whether it approximates CF correctly. Especially given that annotated segmentation masks are available, this should be straightforward to add.

Response:

Yes, the model is trained at the pixel level and our evaluation also include pixel-wise metrics. As per the reviewer's concern, we have also revised the relevant sentences in the manuscript (Line nos 161 to 170) and Table 1 to explicitly state about this clarification.

The values reported in Table 1 - namely, accuracy, F1-score, and ROC-AUC, are all computed at the pixel level using the predicted cloud/non-cloud masks against the ground truth. These metrics directly reflect the per-pixel classification performance, consistent with the model's pixel-wise training shown in Fig. 1. The cloud fraction (CF) metric was also reported to provide a scene-level perspective on the model's behaviour, as CF is often used in climate-related or radiative transfer applications.

We have now added the precision, recall and IoU scores of all the locations in Table 1 in the revised manuscript and provided the confusion matrix for the same in the supplementary as Fig.S2. The revised text is mentioned below for the reviewer's reference.

For each of the locations using TSI, a set of 300 images was selected for that particular location to train a random forest classifier. While the set of images is a representation of different cloud fractions, they also encompass various cloud types, weather conditions, and lighting scenarios of each location. The classifier was configured with 100 trees and a fixed random seed to ensure the reproducibility of results. A train-test split of 80:20 was applied to the dataset, and after training, the model was used to classify cloud and non-cloud pixels of each sky image in the test set.

Each model, trained specifically using images of that location, was used to predict the cloud pixels from the test images corresponding to that location. We computed various performance metrics, including accuracy, F1-score, precision, recall, ROC-AUC score and Intersection over Union (IoU) score to assess the classifier's effectiveness in distinguishing between cloud and non-cloud regions, which is

tabulated in Table 1. The confusion matrix for each location is provided in Fig.S2 of the supplementary, along with the description of each performance metric.

Table.1 shows the performance metrics for each dataset location

Location	***Accuracy***	***Precision***	***Recall***	***F1 Score***	***ROC-AUC Score***	***IoU Score***
Black Forest, Germany	*0.93*	*0.88*	*0.88*	*0.88*	*0.88*	*0.79*
Lamont, Canada	*0.89*	*0.84*	*0.90*	*0.87*	*0.85*	*0.76*
Darwin, Australia	*0.91*	*0.88*	*0.93*	*0.90*	*0.87*	*0.80*
Gadanki, India	*0.94*	*0.85*	*0.92*	*0.88*	*0.90*	*0.79*

In Supplementary:

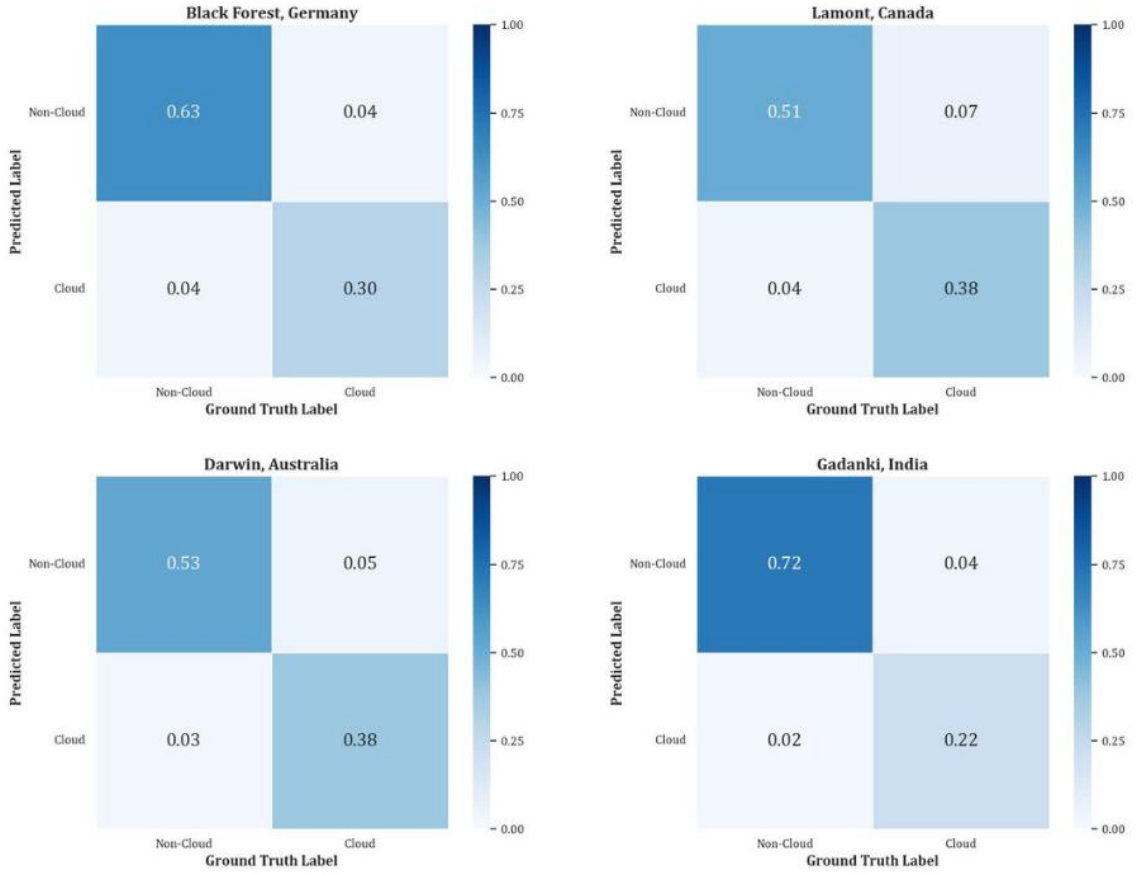

Fig.S2. Confusion matrix obtained from the test set for each location.

---

2. Though the authors mention performance degradation due to sun glare and cirrus clouds, this is illustrated only through a few hand-picked examples. There is no quantification of how prevalent these issues are in the dataset, nor an analysis of how performance varies across such conditions. Similarly, no confusion matrix or class-specific breakdown is presented to identify key failure modes. A more systematic error analysis would strengthen this part.

Response:

We thank the reviewer for this thoughtful suggestion. We agree that a systematic error analysis would strengthen the discussion of model limitations. In our current manuscript, we included representative failure cases (e.g., due to sun glare and cirrus clouds) to illustrate challenging scenarios qualitatively. Moreover, as per your suggestion, we have added additional analysis on the CF errors caused if cirrus clouds or sun glare (figure attached below for reference).

Among the 500 images in the validation set of Merak, 1.6% of the images had cirrus clouds with a mean CF error of 0.14 ± 0.04. About 4.2% of the validation set had sun glare with a mean CF error of 0.12 ± 0.02. We have also updated our manuscript to highlight these errors.

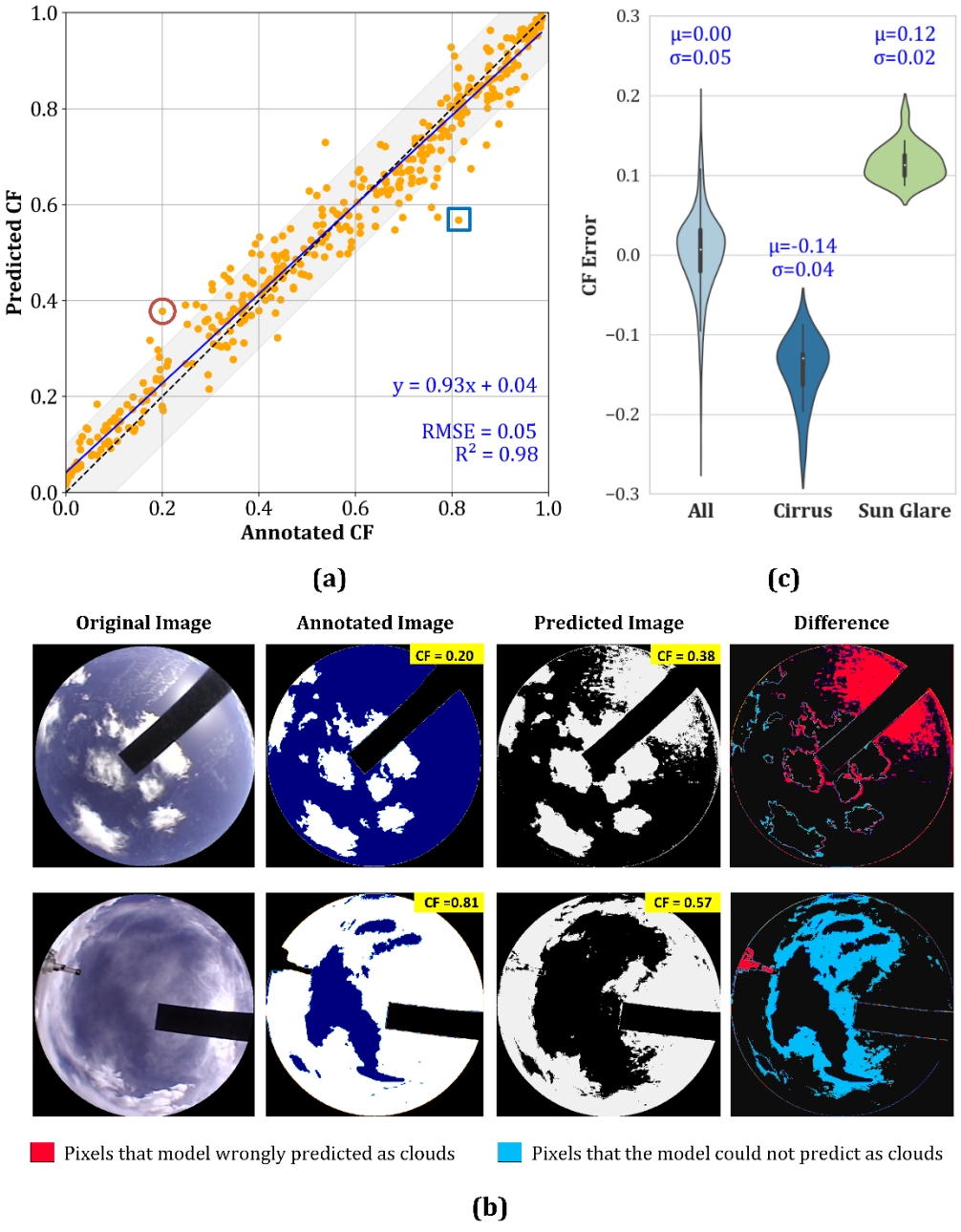

Figure 5: (a) Validation of RF classifier output for images taken at Merak, India (b) Representative failure cases: top row shows overprediction due to sun glare (highlighted by red circle in (a)), and the bottom row shows underprediction caused by cirrus clouds (highlighted by blue square in (a)). Red and

blue pixels in the difference column, indicate misclassified pixels. (c) violin plot that compares CF errors for cirrus and sun glare cases

---

3. The authors highlight that RF is computationally efficient, but there is no measurement of runtime, memory usage, or inference speed. Even a simple runtime comparison on a CPU vs. a lightweight CNN would be informative.

Response:

We thank the reviewer for this helpful observation. The statement that the Random Forest (RF) model is computationally efficient was to highlight a known advantage of RFs as established in prior literature (Mu et al., 2017). Our primary focus was on demonstrating that an RF, when trained with appropriate features, can achieve competitive performance in cloud detection tasks.

However, as per reviewer's suggestion, we performed some inference benchmarks for the RF classifier. We have revised the manuscript with the following information:

To evaluate the computational performance of the proposed model, the RF classifier's inference benchmarks were run on a desktop machine with an Intel Core i7-11700 CPU (8 cores, 16 threads), 16 GB RAM, and no GPU acceleration, running Windows 11 (64-bit). Inference was performed on 280×280-pixel images (~78,400 pixels) with an average runtime of 0.113 seconds per image, a peak memory usage of 41 MB, and an effective processing speed of approximately 800,000 pixels per second. These results reflect the classifier's suitability for real-time, low-power applications without the need for specialized hardware.

….

Reference:

Mu, X., Ting, K. M., and Zhou, Z.-H.: Classification Under Streaming Emerging New Classes: A Solution Using Completely-Random Trees, IEEE Trans Knowl Data Eng, 29, 1605–1618, https://doi.org/10.1109/TKDE.2017.2691702, 2017

Other minor comments:

Line 30: "satellite-based imagers have lower temporal resolutions". The authors ignore the fact that geostationary satellites provide very high temporal resolution (10-minute or better) imagery. This should be acknowledged to give a more balanced view.

Response:

We thank the reviewer for bringing this to our attention. We have revised the statement as follows:

Satellite imagers observe clouds over larger spatial domains (Verma et al., 2018), often with temporal resolutions as good as 10 mins (Huang et al., 2019).

References:

Huang, Y. I., Siems, S., Manton, M., Protat, A., Majewski, L., and Nguyen, H.: Evaluating himawari-8 cloud products using shipborne and CALIPSO observations: Cloud-top height and cloud-top temperature, J Atmos Ocean Technol, 36, 2327–2347, https://doi.org/10.1175/JTECH-D-18-0231.1, 2019.

Verma, S., Rao, P. V. N., Shaeb, H. B. K., Seshasai, M. V. R., and PadmaKumari, B.: Cloud fraction retrieval using data from Indian geostationary satellites and validation, Int J Remote Sens, 39, 7965–7977, https://doi.org/10.1080/01431161.2018.1479792, 2018.

Line 95: "images captured during rain were also removed". Please clarify how rain-contaminated images were identified. Was this done manually or through an automated threshold/filter?

Response:

We thank the reviewer for this important clarification request. The identification of rain-contaminated images was performed manually, based on visible artifacts such as raindrops on the lens, severe blurring, or overall low visibility that typically accompany rain events. These images were visually inspected and excluded during dataset curation to ensure the model was trained only on usable sky conditions.

We have revised the manuscript to explicitly state that the removal process was done through manual inspection.

Line 123: "Random Forest" -> Should be abbreviated as RF.

Response:

The manuscript has been updated with the correction.

Line 137: The sentence stating that RF models are "difficult to interpret" is vague. Please be specific: are the authors referring to the difficulty of tracing individual pixel classifications back to specific trees or features? If so, mention this explicitly.

Response:

We thank the reviewer for this insightful observation. We have revised the sentence as follows for more clarity:

A key limitation of Random Forests is that, due to their ensemble nature, it is difficult to trace individual pixel-level classifications back to specific features or decision paths.

Line 159 (Figure 1 caption): The caption is too terse. I would expect a more informative caption that explains the key steps in the algorithm flowchart.

Response:

We thank the reviewer for bringing this to our attention. We have revised the caption as below, to provide a more detailed and informative caption that clearly explains the data flow, training pipeline, and evaluation steps of the proposed methodology as shown in the flowchart.

Figure Caption:

Fig.1. Workflow of the Random Forest-based cloud detection framework. The input images are pre-processed and annotated to create a master dataset, which is then split into training, validation, and test sets. The Random Forest model is trained with hyperparameter tuning and evaluated on validation data. The trained model generates predicted cloud masks, from which cloud fraction is computed and compared against ground truth for output validation.