Author responses to Referee #1 comments on "Leveraging Machine Learning techniques and SEVIRI data to detect volcanic clouds composed of ash, ice, and SO_2"

Manuscript: egusphere-2026-727

We would like to thank Anonymous Referee #1 for his constructive comments and suggestions, which have improved the manuscript. Please find our replies to each comment below.

This document is structured according to the following color legend:

- Reviewers' comment.
- Author's response.
- Author's changes implemented in the manuscript.

Naranjo et al. report a new algorithm for volcanic cloud detection using a spaceborne visible and infra-red imager. They conclude that their approach improves on current algorithms for clouds that include a mixture of volcanic ash, ice, and SO2. The weakest performance of the new algorithm occurs at cloud edges; an expected result of weaker signal. The methods are good and the results generally supportive, so I am happy to recommend publication with revisions that address the following concerns. Overall, the manuscript was a pleasure to read.

1. Whether or not the performance metrics in Table 5 are "high" should be considered relative to performance of previous cloud detection algorithms. The discussion (line 349) includes references to previous algorithms that fail in the presence of ice mixtures. The authors could show quantitative support for their main conclusion by running those algorithms over their validation case studies in order to make a direct comparison.

Done. Thank you. Table 5 in the revised manuscript now includes a comparison of the evaluation metrics obtained using the BTD[11–12] method, the raw NN plume mask, and the filtered NN plume mask.

L349-351: This section presents the detection results obtained for these events, which are summarized in Table 5. Table 5 presents the performance metrics derived from the filtered NN plume mask, together with those obtained from the raw NN plume mask and the BTD method for comparison.

Table 5. Overall performance metrics for the complete image sequences of the three validation events obtained using the filtered NN plume mask. Metrics derived from the BTD method and the raw NN plume mask are also presented for comparison.

Event	Method	Balanced Accuracy	Precision	Recall	F1-Score
1: 22 Feb 2021	BTD[10.8-12.0] < 0.0	0.4856	0.0049	0.0116	0.0069
	Raw NN plume mask	0.7183	0.3663	0.4490	0.4035
	Filtered NN plume mask	**0.8293**	**0.9234**	**0.6595**	**0.7695**

	BTD[10.8-12.0] < 0.0	0.4549	0.0007	0.0167	0.0014
2: 4 Mar 2021	Raw NN plume mask	0.6532	0.1514	0.3152	0.2046
	Filtered NN plume mask	**0.7424**	**0.9601**	**0.4850**	**0.6444**
	BTD[10.8-12.0] < 0.0	0.5405	0.0965	0.0908	0.0935
3: 4 Aug 2024	Raw NN plume mask	0.7011	0.3816	0.4100	0.3953
	Filtered NN plume mask	**0.9196**	**0.7812**	**0.8419**	**0.8104**

2. There are deficiencies, and possibly some error, in the reporting of algorithm metrics that I can frame around Figure 8.

Thank you. Figure 8 was improved and the identified errors were corrected. For example, Figures 8d–f were originally presenting false negatives, although the label incorrectly indicated false positives. In addition, in some parts of the text, the false positive metric was mistakenly interchanged with the false negative metric. These inconsistencies throughout the text have now been corrected.

The authors show metrics for the validation case studies, which has highly imbalanced classes, but not for the test subset of the labelled data which has been sample to improve balance. The latter metrics, on the balanced test set, should also be shown. It will allow readers to know how much value comes from the NN verse the post-processing.

Done. Thank you. The evaluation metrics for the raw NN output (before the post-processing step) are presented now in Table 5 in the revised manuscript. Additionally, in the revised manuscript, we used the balanced accuracy metric, which is more appropriate for imbalanced datasets such as the validation case studies (see Table 5)

L170-174: A trade-off was therefore established between achieving a training accuracy of 96.0% and maintaining an average inference time of approximately 1.5 seconds per classification. The final selected hyperparameters combination is reported in Table 3 (Section 3.2 Training phase). Using these hyperparameters, the evaluation metrics obtained for the testing set were 95.0% accuracy, 98.6% precision, and 91.3% recall (see Section 3.5 for the definition of these metrics).

Optionally, the authors could include the metrics for each set of hyper-parameters in a supplement.

Done. Thank you. The hyperparameter tuning results can now be found in the file results_hyperparameter_tuning_sorted.csv in the supplementary materials.

The confusion matrices ought to be rotated (with "Observed" on the horizontal), normalized by the total number of sample (not the marginal totals), and the calculation checked: the counts have been normalized by the total observed in each class, so the resulting true-positive-rate (currently bottom right corner) should equal the recall in table 5.

Thank you. Your suggestion was followed. In the new Figure 8, the confusion matrices were rotated, with the "Observed" labels presented along the horizontal axis and the "Predicted" labels along the vertical axis. This new organization of the confusion matrix corresponds to the layout presented in Fawcett (2006), see the figure below.

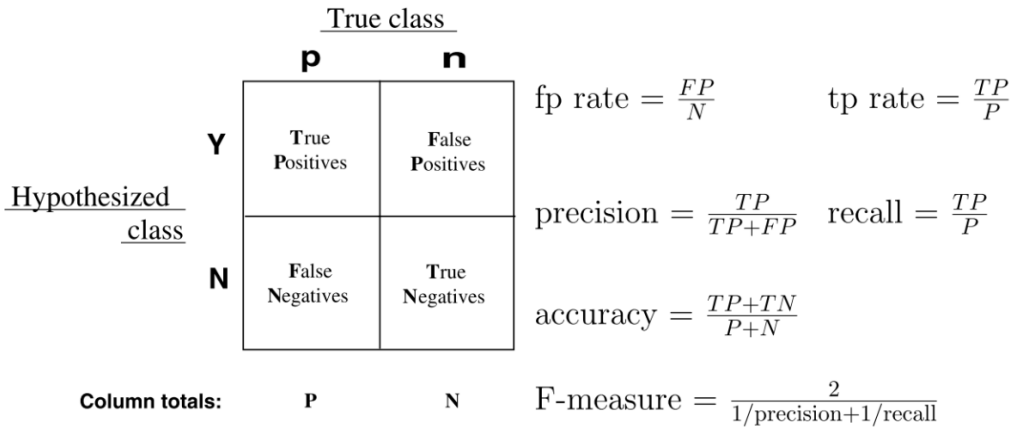

Fig. 1. Confusion matrix and common performance metrics calculated from it.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognit Lett, 27, 861–874, https://doi.org/10.1016/J.PATREC.2005.10.010, 2006.

In addition, the confusion matrices are no longer normalized, and the total number of pixels is now displayed. We believe that this representation facilitates the calculation and verification of the evaluation metrics (recall, precision, etc.) presented in Table 5.

Additionally, in the revised manuscript, we revised the evaluation metrics and now report the recall, precision, and F1-score specifically for the VC class. In the original version of the manuscript, macro-averaged metrics were presented, which compute the average performance across both classes. However, because the validation case studies are highly imbalanced, the use of macro-averaged metrics may lead to misleading interpretations of the results.

The authors need to say what threshold they used for class assignment and why.

Thank you. This criterion is now stated and discussed at the end of Section 3.2, Training Phase.

L209-211: This symmetric and bimodal distribution indicates that only a small number of ambiguous probability values are present, which is a desirable characteristic of a well-calibrated classifier. Based on these results, we decided to classify a pixel as VC when the calibrated probability is higher than 0.8.

3. In section 4.4, the authors struggle (as do I!) to reason about the separate utility of precision and recall. Given the highly unbalanced class composition in the dataset, and the noted application to aviation safety, it may be better to simply focus on recall (the rate of false negatives).

Thank you. In the revised manuscript, we focused primarily on the recall metric, while also discussing precision. In addition, the discussion of the evaluation metrics was improved to facilitate a clearer understanding of the results.

4. Encountering the methods and results of a final analysis in the discussion is surprising. Consider rearranging.

Thank you. A new subsection entitled 3.6. Retrieval Comparison was included, to which the description of this comparison was relocated.

5. I encourage the authors to publish any software or scripts developed for this manuscript as a code supplement. I also encourage the authors to plan for release of the labelled dataset after a sufficient period of embargo; choose a data archive that makes the labelled dataset citable and only the metadata public and discoverable. At some later date it will then be easy to make the labelled data itself open, which would be a great contribution to future research.

Done. Thank you. The balanced dataset used in this study was published, and a Jupyter Notebook for reading the dataset was made available in a GitLab repository.

Dataset:

Naranjo, C., Guerrieri, L., Corradini, S., Picchiani, M., Merucci, L., & Stelitano, D. (2026). Balanced Dataset of SEVIRI Observations for the Detection of Volcanic Clouds Composed of Ash, Ice, and SO2 (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.20313629

Repository GitLab:

https://gitlab.rm.ingv.it/camilo.naranjo/etna_volcanicclouds_2020-2022_dataset