Author responses to Andrew T. Prata 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 Andrew T. Prata 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.

General comments

In this paper, Naranjo et al. propose a new volcanic cloud detection algorithm using a neural network (NN) applied to eruptions from Mt Etna. The main issue they tackle is volcanic cloud detection when the volcanic clouds do not exhibit a 'traditional' volcanic ash signal (i.e. a negative 11-12 BTD). They show that this issue is common for a series of Etna eruptions from 2020-2022 and 2024. The authors rely on manual labelling, primarily informed by Ash RGB composites, to derive training data and build a balanced, 50/50 Volcanic Cloud (VC) and non-Volcanic Cloud (NVC), labelled dataset for training and testing. The authors found very good results when validating their algorithm against three Mt Etna case studies that the algorithm had not seen in the training phase (accuracy=99%, precision > 89% and recall >74%). The authors also find that performance degrades over time as the VCs become more dilute and detection sensitivity becomes challenging. Interestingly, the authors find that the false positive rate is lowest for the most complex case which they argue might be because the human operator (manual labeller) tends to be more conservative in these situations (drawing a smaller volcanic cloud boundary) and therefore dilute pixels (where the NN may struggle) are less likely to be labelled as VC. Neural networks are becoming increasingly useful in feature detection in satellite imagery and I am pleased to see this method being applied to the complex task of classifying volcanic clouds composed of ash, ice and SO2. I certainly find the contribution novel and that it addresses a longstanding and complex problem - detection of hazardous volcanic clouds that are not dominated by an ash particle spectral signature. The theme of this review is essentially to ask for further details. In that respect my comments are mostly minor, depending on how the authors choose to address them. I don't believe the analysis needs to be redone - just more details on how it was performed are required. I would also comment that providing the manually labelled dataset to the research community would be very beneficial. This would allow researchers to design and compare new algorithms based on this valuable training dataset that has been put together.

Thank you. The balanced dataset 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

Specific comments

Training dataset

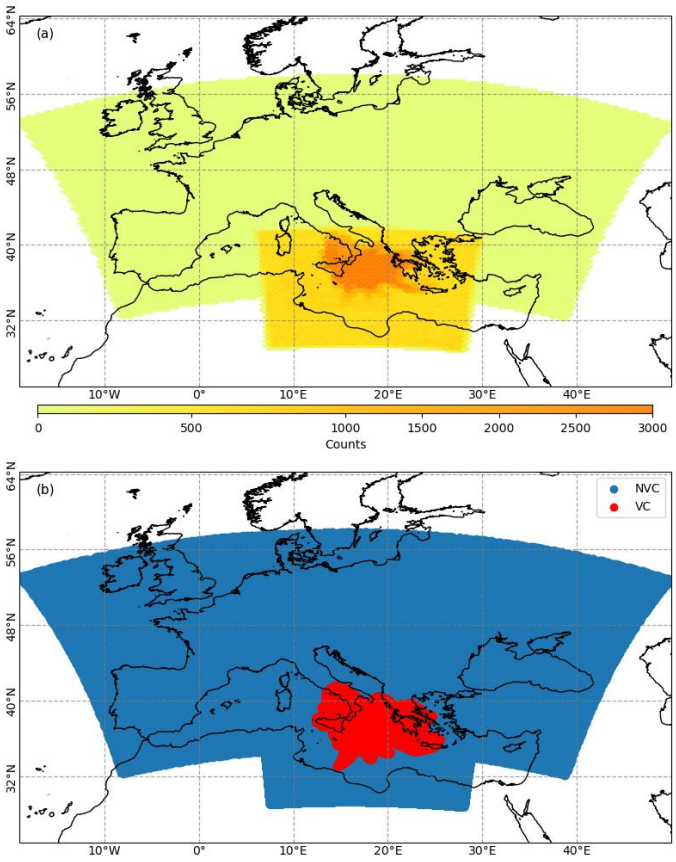
It would be useful to have some more detail on how the training dataset was constructed. I'm also interested in understanding how the data were randomised. Can the authors provide the spatial region that was used for the analysis (i.e. latitude/longitude box) and how many pixels are contained within this region? Was there any consideration of balancing the non-VC part of the dataset by defining other non-VC classes? I'm wondering if the non-VC pixels were dominated by some other non-VC class (e.g. is it mostly clear-sky pixels?). Similarly, in time, were the non-VC pixels only sampled during eruptions of Mt Etna? These questions have relevance for understanding any bias that might be in the non-VC labelling e.g. If you only sample non-VC pixels in the vicinity of Etna, then the technique might not generalise to other regions of the world (Indonesia, Iceland etc). Similarly, if pixels are only selected during Etna eruptions then it is possible that you may have a temporal bias in your non-VC class but only if the Etna eruptions don't adequately cover all seasons.

Thank you. Along with the publication of the balanced dataset, we added a document providing a brief description of the dataset. This document includes additional information regarding the spatial and temporal coverage of the dataset.

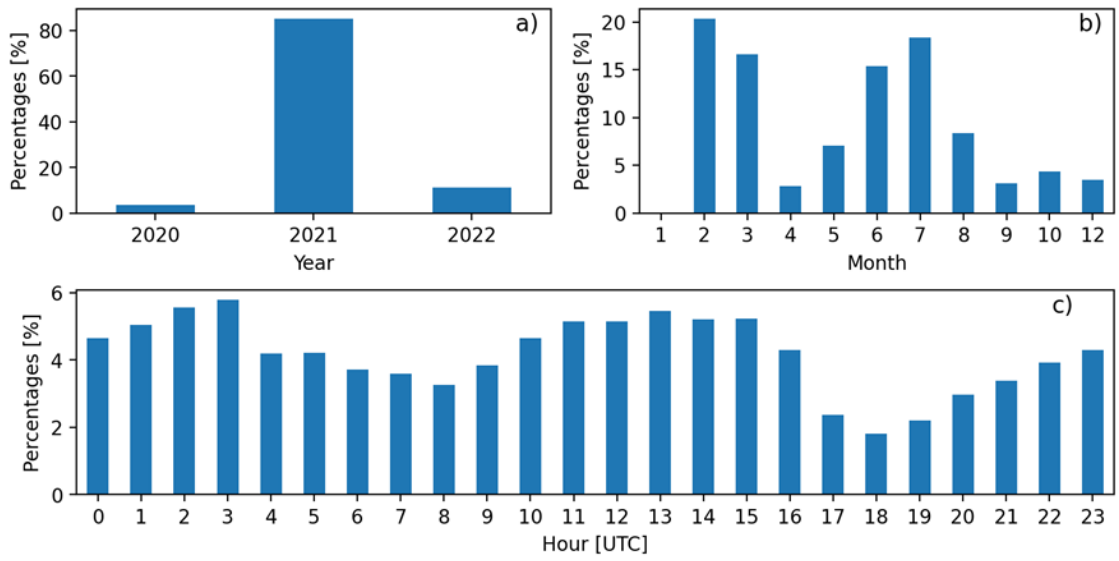
In addition, more details regarding the construction of the balanced dataset were included in Section 2.3 Dataset, of the revised manuscript.

Since the NVC class refers to non-volcanic features (e.g. meteorological clouds, land surfaces, and sea surfaces), it was not possible to precisely determine how these pixels were sampled during the random sampling process. An interesting direction for future work would be to use all NVC pixels that were discarded during the random sampling procedure. Moreover, the NVC pixels were sampled over the entire available area and not only in the vicinity of Mt Etna.

Regarding the temporal and spatial coverage, see the figures below, which were included in the dataset overview published on Zenodo.



Spatial coverage of the dataset: (a) distribution of the total number of pixels; (b) spatial distribution of the dataset for each class, with VC shown in red and NVC shown in blue.



Temporal coverage analysis of the dataset: (a) yearly distribution, (b) monthly distribution, and (c) hourly distribution.

Selection of BT and BTD features

I'm interested in understanding how these combinations were chosen. Some are clear (e.g. 11-12 BTD for ash presence) but others are new and interesting (e.g. 11 - 9.7). One could construct several more BTD combinations (e.g. 7.3 - 6.2 may contain a UTLS SO2 signal), but I'm wondering if any statistical technique was used to choose the BTD combinations to begin with? Or was it simply the approach of just pinning all the differences to the 11 µm channel?

Thank you. It is an interesting question, as most of the time in the development of a Neural Network model is spent on this step. We conducted several experiments using combinations of fixed hyperparameters while varying the features included in the dataset. Some results are shown in the table below.

Our first approach was to use only the seven thermal infrared channels. Although the accuracy results were good, we noted that the model was slow. Based on these results, we decided to follow a strategy of aggregating band combinations that are useful for human interpretation of satellite images, using RGB composite recipes. Thus, by considering the BTDs commonly used in RGB composites, we decided to aggregate all BTD combinations.

Regarding the inclusion of the UTLS (BTD[7.3 – 6.2]), this was interesting because we considered including this band instead of BTD[108–062] and BTD[108–073]. However, no significant differences in accuracy were observed, as shown in the table below. Therefore, to maintain consistency and standardization in the selected features, we decided to include only the BTDs proposed in this study.

Nevertheless, we believe that the NN model could be further optimized by selecting more or fewer features according to their importance. However, this process is time-consuming and requires substantial computational resources.

Features	mean_fit_time [s]	mean_score_time [s]	mean_test_score	mean_train_score
'WV_062', 'WV_073', 'IR_087', 'IR_097', 'IR_108', 'IR_120', 'IR_134',	1875.168054	2.440265	0.956911	0.958055
'WV_062', 'WV_073', 'IR_087', 'IR_097', 'IR_108', 'IR_120', 'IR_134', 'BTD[108-062]', 'BTD[108-073]', 'BTD[108-087]', 'BTD[108-097]', 'BTD[108-120]',	**1298.729537**	**1.514791632**	**0.959605672**	**0.96117258**

'BTD[108-134]',				
'WV_062', 'WV_073', 'IR_087', 'IR_097', 'IR_108', 'IR_120', 'IR_134', 'BTD[062-073]' 'BTD[108-087]', 'BTD[108-097]', 'BTD[108-120]', 'BTD[108-134]',	1487.968672	1.385069	0.95961	0.961306

Clarification around ML methods

Some choices are simply stated without much motivation or reasoning. I wonder if this is because the choices are standard practice in ML research (which I am not an expert in)? In my line-by-line comments I have raised questions where I don't fully understand what was done. Any attempt to address these questions will help improve the clarity of the paper.

Done. Thank you. In your Line-by-line comments we addressed all the questions and the corresponding information was added in the revised manuscript.

Plume tracking

The plume tracking algorithm looks really effective. However, I'm unsure how the analysis presented in Fig. 4 is related to the analyses presented later in Figs. 5, 6 and 7. For example, how is the "Enclosing" purple line on Figs. 5, 6, 7 related to the expanding circular radius in Fig. 4? Also, how do you prevent the circular region in the plume tracking from growing very large? If the circular region is expanded at every time step, then for long lasting eruptions (>12 h) like the Fig. 6 (4 March 2021) for example, what prevents the region from continuously growing (thus potentially introducing more false positives)? Further, if this region is not limited, then could this be partly the reason for your increase in false positives with time? I.e. the plume tracking radius is becoming very large and is introducing noise from other regions far away from the volcanic cloud?

Thank you. As was explained in the Figures comments, the purple enclosed polygon was included just for visualization purposes, following the same approach adopted by the Volcanic Ash Advisory Centers (VAACs) for reporting volcanic clouds and by the VOLcanic Cloud Analysis Toolkit (VOLCAT) described in Pavolonis et al. (2018). However, to improve clarity and avoid potential misunderstandings, it has been removed from the figures 5, 6 and 7 in the revised manuscript.

Additionally, some comments regarding the plume tracking algorithm were added to the manuscript. In this work, the tracking algorithm does not include a control to prevent excessive growth of the circular region. The presented version was designed to automatically detect

volcanic clouds from a given sequence of geostationary observations. We agree that this limitation may become problematic when tracking volcanic clouds over long time periods, for example in near-real-time applications. Under such circumstances, a more sophisticated algorithm should be developed.

Regarding the false positives, we first clarify that, in some parts of the text, the false positive metric was mistakenly interchanged with the false negative metric. These inconsistencies have now been corrected throughout the manuscript. Returning to your last comment, we agree that an excessively large plume tracking radius could introduce noise from regions far from the volcanic cloud. However, this behaviour was not observed in the validation experiments presented in this study. While, we think the increase in false negatives is unlikely to be related to plume tracking radius.

L263-267: Unfortunately, in the presented version of the post-processing algorithm, not all "noisy" pixels can be effectively removed through the application of the NL means filter. In addition, the tracking algorithm does not include a control to prevent excessive growth of the circular region, which may become problematic during near-real-time operations when tracking volcanic clouds over long time periods. Under such circumstances, a more sophisticated post-processing algorithm could be developed.

Evaluation metrics

I know the metrics reported are standard in ML literature but I think it would be beneficial to include their mathematical definitions when describing how these metrics were implemented as there are several pre- and post-processing steps. I also think the confusion matrix results should potentially be reported in the abstract or at least be more upfront alongside the accuracy, precision and recall. The analysis presented in Fig. 8 and the discussion around it are very useful and insightful. Here we see how the algorithm's performance degrades over time due to the volcanic cloud signal becoming "diluted" as the authors hypothesise. But I also find it useful to know that the model only classifies a VC when it actually is a VC 66%, 49% and 84% of the time for the three validation events (assuming I'm interpreting the confusion matrix plots correctly). I think these numbers are more insightful than, say, the accuracy metric (e.g. 99% accuracy sounds like you have a 'perfect' model).

Thank you. In the revised manuscript, a new section entitled 3.5. Evaluation Metrics was included to define all the metrics used for the validation process. In addition, more information describing how these metrics were applied was added.

L269-270 : Finally, the results and evaluation metrics presented in the Results section correspond to the filtered plume masks obtained after the application of this post-processing step.

L305-307: It is worth noting that the accuracy metric was used for the training and testing datasets, whereas balanced accuracy was used for the validation dataset. This choice was made because the validation dataset is highly imbalanced, making balanced accuracy metric more appropriate for this type of dataset.

Section: 3.5. Evaluation metrics

Regarding the confusion matrix, the recall and precision information, these metrics were included and made more prominent in the Abstract, Discussion, and Conclusions sections.

Taking into account the comments from Reviewer #1, the configuration of the confusion matrix in Figure 8 was modified, and the performance metrics presented in Table 5 were carefully revised. In the revised manuscript, the reported metrics provide a clearer and more representative evaluation of the actual performance of the NN model.

Mt Etna VC compositions and mass loading retrievals

The mass loading comparison is useful as it demonstrates that while the NN may miss the diluted plume pixels, it is not too serious in the context of total mass loading estimates because the diluted pixels do not contribute much to the total mass per unit time. However, I'm not sure I understand how the individual mass loading time-series were calculated for the NN masks if the algorithm simply classifies pixels as Volcanic Cloud (VC) or non-VC. If the pixel is "VC", how do you separate the mass loading into ice, ash and SO2? I think it would be useful to define what you mean by VC. E.g. Are volcanic clouds = ash + SO2 + ice, or could they be either ash, SO2 or ice or something else? Also, a component that hasn't been mentioned is sulphate. Have the authors considered the influence of sulphate aerosols on the RGB signals and BTD combinations they're using? E.g. Sellitto and Legras (2016): https://amt.copernicus.org/articles/9/115/2016/

Thank you. A new subsection entitled 3.6. Retrieval Comparison was included to provide a clearer explanation of how the mass loading comparison was performed. See the L354 comment in your Line-by-line comments for more details.

We also specified the meaning of the VC and NVC classes.

L126-128: The VC class refers to volcanic cloud pixels containing ash, SO2, ice, or any combination of these components, whereas the NVC class refers to pixels associated with meteorological clouds, land surfaces, sea surfaces, and other non-volcanic features.

Thank you for mentioning sulphate aerosols as an important component and for providing the reference. In this study, we did not consider the influence of sulphate aerosols on the RGB signals and BTD combinations; however, we believe this represents an interesting topic for future studies. A possible starting point would be to determine in which parts of the volcanic clouds analysed in this study sulphate aerosols are present.

Conclusions/future work

Based on results and analysis in the paper, it appears to me that the authors have a very good algorithm for detecting ash+ice+SO2 clouds from Mt Etna. This kind of analysis would really benefit VAACs that deal with these types of clouds routinely e.g. the Darwin VAAC routinely see volcanic clouds rich with ice and SO2. I think an area for future work could be to see whether or

not the algorithm can characterise ash/ice/SO2 clouds from other eruption case studies (in different atmospheres e.g. Russian/Alaskan/Icelandic volcanoes at high latitudes or Indonesian volcanoes in the tropics). I agree that addressing the diluted parts of the plume is an area for improvement. However, I think it would also be valuable to understand if this method/algorithm would generalise to other geostationary sensors and/or other eruption case studies. Or would a new algorithm (maybe with the same network architecture and input features) need to be re-trained on the relevant dataset (e.g. Himawari/GOES)? Some comment or recommendation on this aspect of the research would be helpful.

Done. Thank you. We added some comments about this aspect in the Conclusions section.

L593-601: Overall, the results are encouraging and provide valuable insight into the development of a near-real-time and automatic detection system for volcanic clouds, which is highly desirable for aviation safety and could be useful for the VAACs that deal routinely with volcanic clouds rich with ice and SO2.

An important question for future studies is to determine whether it is feasible to apply transfer learning using the same NN model to extend its application to new generations of geostationary sensors (e.g. Meteosat-12 FCI, Himawari AHI, and GOES ABI) and to additional eruption case studies worldwide. Addressing this question would require experiments to evaluate the performance of the NN model using data acquired from different sensors and volcanic eruptions occurring at different geographical locations. This approach could eventually lead to the training of a new NN model based on the same architecture adopted in this work, but using representative datasets derived from the new case studies and satellite sensors.

Line-by-line comments

- Title: I don't think you need to capitalise "Machine Learning". Just use "machine learning".

Done. Thank you.

Leveraging machine learning techniques and SEVIRI data to detect volcanic clouds composed of ash, ice, and SO2

- L12: "Etna volcano" -> "Mount Etna volcano (Italy)".

Done. Thank you.

… instrument. A dataset of 1.259 SEVIRI images related to Mount Etna volcano (Italy) eruptions spanning from 2020 to 2022 …

- L14: "BTD's" -> "BTDs"

Done. Thank you.

… including thermal infrared channels and brightness temperature differences (BTDs). The model was validated on three eruptive …

- L30-31: I understand the concept but for readers not familiar with the research area I suggest being more explicit by revising to: "Specifically, ash absorbs more energy around the 11 µm band than the 12 µm band, whereas water and ice absorb more energy around the 12 µm band than the 11 µm band. "

Done. Thank you.

Specifically, ash absorbs more energy in the 11 µm band than in the 12 µm band, whereas water and ice absorb more energy in the 12 µm band than in the 11 µm band.

- L57: Again does "Machine Learning" need to be capitalised? Why not just "machine learning"?

Done. Thank you.

- L83: "SEVIRI instrument" -> "The SEVIRI instrument"

Done. Thank you.

… this work. The SEVIRI instrument can produce an image of the Earth's full disk every 15 minutes in 12 different spectral …

- L86: "All the SEVIRI data used in this work was…" -> All the SEVIRI data used in this work were…"

Done. Thank you.

… wavelengths are shown in Table 1. All the SEVIRI data used in this work were acquired in near real-time using the …

- L91: When describing the natural color composite can you provide which SEVIRI channels (wavelengths) are assigned to R, G, B? Same goes for the Ash RGB. I know it can be found in the EUMETSAT guide but it would be useful to have them at hand when reading the paper and it's easy to insert this information.

Done. Thank you.

The natural color composite, created by assigning the 0.635, 0.81, and 1.64 µm channels to the red, green, and blue color beams, respectively, is shown in Figure 1(a) …

- L93-95: Ash RGB - can you state how it's constructed (see comment above)? I know it's not simply single channel BTs assigned to R, G, B.

Done. Thank you.

By contrast, the volcanic cloud shown in Figure 1(b) is well distinguished in the Ash RGB composite (Ash RGB Quick Guide | EUMETSAT - User Portal, 2025), which uses the thermal infrared channel. In this latter figure, generated by visualizing the channel differences IR12.0 – IR10.8 and IR10.8 – IR8.7 in the red and green color beams, respectively, and the IR10.8 channel in the blue beam, the constituents of the volcanic cloud are clearly identifiable …

- L120-123: Pre-processing of BTs and BTDs before input into the model - I see later on you state: "Additionally, the data were standardized by removing the mean and scaling to unit variance". I think it would be useful to have this information at this point in the paper. Also, how is this actually done in practice? How is the mean computed? Is it the mean of all pixels considered (VC+nonVC)? Or do you need to group by category and then take the mean?

Thank you for this comment. In this case, we decided to keep the information about the pre-processing step in its original position.

We can see your point about the original data (BTs and BTDs) are pre-processed before input into the model. However, the standardization step is a standard and common normalization procedure in many machine learning techniques. Therefore, the input data are still the original BTs and BTDs, although they are transformed before input into the model as part of the machine learning pipeline.

Regarding the standardization step, it is first applied independently to each feature in the Training set, in this case each BT or BTD, including data from both classes (VC and NVC). The mean and standard deviation are calculated for each feature. Subsequently, each feature in the Test set is standardized using the mean and standard deviation previously computed from the Training set.

In practice, the mean and standard deviation computed from the Training set must be stored, as they are required each time the model is used to classify a new image. That mean and standard deviation are important because they contain the statistical information of the Training set.

Finally, we have added more information about the standardization step in the manuscript. Please see the text below:

L154-157: Additionally, the data for each feature was independently standardized by removing the mean and scaling to unit variance, the process is also known as Z-score normalization. In this standardization step, data from both classes (VC and NVC) were included to calculate the mean and standard deviation for each feature. Standardization is an important requirement for NN models, as it generally improves their performance (Pinheiro et al., 2025).

Pinheiro, J. M. H., Oliveira, S. V. B. de, Silva, T. H. S., Saraiva, P. A. R., Souza, E. F. de, Godoy, R. v., Ambrosio, L. A., and Becker, M.: The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks, IEEE Access, 13, 199903–199931, https://doi.org/10.1109/ACCESS.2025.3635541, 2025.

- L140: Parameter optimisation: Is this what is typically referred to in the literature as "hyper-parameter tuning"? If so, I wonder if similar language should be used. If not, how is this method different from hyper-parameter tuning?

Thank you. In the literature, terms such as parameterization, parameter optimization, training optimization, and hyperparameter tuning are all used. These terms refer to the process of adjusting the model hyperparameters to minimize the cost function during the training phase.

To improve clarity and adopt the most commonly used terminology, we followed your recommendation. We changed the title of Section 3.1 to Hyperparameters and consistently used the terms hyperparameter and hyperparameter tuning throughout the text.

3.1. Hyperparameter tuning phase

To determine the most suitable configuration for the NN model, a hyperparameter tuning phase was carried out. During this step, the hyperparameters listed in Table 2 were optimised using the search space specified in the table's right column. This process combined exhaustive search with a 5-fold cross-validation strategy (k = 5) (Ojala and Garriga, 2010), evaluating all possible combinations within the defined optimisation space. Moreover, the dataset was split into 80% for the training set and 20% for the test set.

- L144: Training/test split - Can the authors be more explicit here e.g. "Moreover, the balanced dataset (50% VC, 50% NVC, totalling 2.2 million pixels) was split…". Also was the train/test split randomised? I see in the validation section you mention that the model hadn't seen some eruption cases. It would be useful to state this at this stage of the paper. It would be interesting to know if the training pixels were sampled evenly across the different Etna eruptions. I've seen 80/20 used many times but wonder if there is any previous literature that indicates the 80/20 rule is best practice for satellite pixel classification?

Thank you. We have added further details regarding the balanced dataset, the validation dataset, and the train/test split. First, we included a general description of the data splitting procedure in Section 3, Model Development:

Here, it is worth defining the convention adopted for the datasets used in the development and validation of the model. For the hyperparameter tuning and training phases, the balanced dataset was divided into training and testing sets. The training set was used to fit the model and support the hyperparameter tuning process, whereas the testing set was used to evaluate the generalization performance of the model and assist in its calibration. In addition, a validation dataset, composed of data from three eruptions not previously seen by the model, was used for

independent validation. These data were not included in the balanced dataset. Further details of the Neural Network model are provided in the following subsections.

Then, in Section 3.1, Hyperparameters, we added further specific details, as recommended:

Moreover, the balanced dataset (50% VC and 50% NVC, totalling 2.2 million pixels) was split into 80% for the training set and 20% for the testing set, in accordance with the proportions commonly employed in the literature (Sun et al., 2022). The split was performed using random sampling with a stratified strategy. This stratified approach ensures that the training and testing sets contain approximately the same proportion of samples from each class (VC and NVC) as the balanced dataset.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W. H., Ma, X., Rao, Y., Bednar, J. A., Tan, A., Wang, J., Purushotham, S., Gill, T. E., Chastang, J., Howard, D., Holt, B., … John, A. (2022). A review of Earth Artificial Intelligence. Computers & Geosciences, 159, 105034. https://doi.org/10.1016/J.CAGEO.2022.105034

About your last two questions:

1. Yes, the balanced dataset splitting process was performed using random sampling across the different Mount Etna eruptions included in the balanced dataset.

2. In the literature, 70/30, 80/20, and 90/10 are the most commonly used proportions for dataset splitting. However, there is no standard best practice. In fact, Sun et al. (2022) state that there is no fixed optimal proportion for dataset allocation, as it largely depends on the characteristics of each study.
   Before selecting the final proportion, we conducted some experiments using these three commonly splitting proportions, and the corresponding results are presented below. To ensure a fair comparison between the different splitting strategies, we report the accuracy obtained during the hyperparameter tunning phase using a fixed set of hyperparameters: three hidden layers, 60 neurons per layers, and both the learning rate and L2 regularization coefficient set to 1×10^{-3}. It can be noted that, for our balanced dataset, there are no significant differences in the accuracy results among the different splitting proportions. Ultimately, we selected the 80/20 split as an intermediate solution. In addition, since the testing set was subsequently used for model calibration, we considered 20% to represent a reasonable trade-off between training and testing data allocation.

Split	**70/30**	**80/20**	**90/10**
Test accuracy [%]	96.12	96.07	96.21

- L145: “both accuracy and the average time required for the NN” - It would be useful to have some numbers to support this statement. For example, were there any settings combinations that had a high accuracy but weren’t used because the run took too long? And how long is “too long”?

Thank you. We have added the numerical criteria used to select the final hyperparameter configuration.

During the hyperparameter tuning phase, both the classification accuracy and the average time required for the neural network to perform a single classification were considered. 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 complete hyperparameter tuning results are provided in the supplementary materials, while the final selected hyperparameter combination is reported in Table 3 (Section 3.2, Training phase).

Regarding your question, we evaluated 504 hyperparameter combinations. Below, we present six representative combinations as examples. Based on the results shown below, it can be noted that the highest accuracy (mean_test_score) is achieved by the configuration with six hidden layers and 100 neurons per layer. However, the corresponding training time (mean_fit_time) is approximately 7087 s (about 2 hours), while the average inference time required to classify a single pixel is approximately 14 s.

Since SEVIRI data contain a large number of pixels (we usually work with an area of interest of approximately 374 × 259 pixels around Mount Etna), the average inference time per pixel becomes an important factor. Therefore, a trade-off between computational efficiency and classification accuracy was considered. Based on this criterion, we selected the hyperparameter combination shown in the fifth row of table bellow (highlighted in bold), which provides an average inference time of approximately 1.5 s together with an accuracy of 96%.

The complete hyperparameter tuning results can be found in the supplementary materials. From the complete ranking, it can be observed that the selected hyperparameter combination was ranked in 42 position. However, this combination represented the best trade-off between computational efficiency and classification performance, as it was the first combination to provide a relatively low inference time while maintaining an accuracy comparable to that of the top-ranked combinations.

mean_fit_time [s]	mean_score_time [s]	L2	hidden_layer_sizes	learning_rate	mean_test_score	mean_train_score
7087.334709	14.0299758	0.001	[100, 100, 100, 100, 100, 100]	0.001	0.967935577	0.971546503
13369.38818	19.08515329	0.001	[100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100]	0.001	0.967707613	0.971066468
6026.985873	10.94196267	0.01	[80, 80, 80, 80, 80, 80, 80, 80]	0.001	0.96309314	0.964891218
3938.885944	5.086182785	0.01	[60, 60, 60, 60, 60]	0.001	0.961678533	0.963432857
1298.729537	**1.514791632**	**0.001**	**[60, 60, 60]**	**0.001**	**0.959605672**	**0.96117258**
156.2888203	0.40933156	0.1	[15, 15, 15, 15, 15, 15]	0.1	0.677308678	0.677295694

It is important to note that these results depend on the hardware specifications of the computer on which the exhaustive search was carried out. In our case, the software was executed on a server with the following specifications:

- CPU: Model Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz, Core(s) per socket: 28

- Memory RAM: 130 GB

- L155: Formatting issue with this line.

Done. Thank you.

- L156: Activation function and optimisation algorithm - are these standard choices? Again, is there any literature you can refer to justify their use over other potential options?

Thank you. We have added further details regarding the criteria used for the selection of the ReLU activation function and the Adam optimisation algorithm, both of which are widely recognized as standard choices in the training of neural network models.

The activation function used was the Rectified Linear Unit (ReLU), which is the standard activation function for NN models based on MLPs (Braga-Neto, 2020; Yang, 2019). For the optimisation algorithm, the Adaptive Moment Estimation (Adam) method was applied, due to its computational efficiency and robustness in the training of models using large datasets (Kingma and Ba, 2014).

- L158: Can you expand upon what the "early stopping strategy" is?

Done. Thank you.

Finally, the maximum number of training iterations was set to 300 and an early stopping strategy was implemented, whereby the training process was terminated when the model performance no longer improved between successive iterations.

- L159-160: "the NN model was calibrated so that the predicted class probabilities correspond to the distribution of the observed pixels in the dataset." - Can you expand on this. What is done mathematically to "calibrate" the NN model? Are you applying transformations to the input BT and BTD data? More detail is needed here to understand the method and how Fig. 3 is generated.

Done. Thank you. For the calibration of the model we do not apply any transformations to the input BT and BTD data; instead, we use the testing set for this purpose. In practice, an additional calibrator model is trained to map the raw output probability of the NN model to a calibrated probability. More detailed information can be found in the manuscript now.

This step is optional, but for us is important because we were interested not only in predicting the class (VC or NVC), but also the associated probability. The probability could be useful when extending the application of the model to other sensors (i.e. FCI-MTG), where the associated probability can act as a detection threshold that can be adjusted.

In addition, the NN model was calibrated. When the model performs a classification, the output corresponds to the probability associated with the respective class, with values close to zero indicating the NVC class and values close to one indicating the VC class. However, the raw output probabilities are not inherently calibrated and may provide unreliable estimates of the true class probabilities. For this reason, a calibration procedure was applied so that the predicted probabilities could be interpreted as the model's confidence that a given pixel belongs to a specific class. The model's confidence specifically depends on the dataset on which it was trained, which in our case corresponds to the balanced dataset. For the calibration the testing set (20% of the balanced dataset) was used to train an additional calibrator model. This calibrator learned to map the raw output probabilities (p_{raw}) of the NN model to a calibrated probability (P_{cal}). In practice, the calibrator predicts the conditional event probability P(prediction = VC | p_{raw}). Therefore, the calibrated probabilities can be interpreted as confidence level.

The calibration curve for the trained NN model is shown in Figure 3a, where the y axis represents the observed proportion of VC class pixels in the testing set, whereas the x axis represents the predicted probabilities for these pixels. The calibration curve was generated by binning predicted probabilities and then plotting the mean predicted probability in each bin against the observed proportion of VC class pixels (Bröcker and Smith, 2007).

As shown in Figure 3a, the NN model exhibits a slight tendency to overestimate low probabilities and underestimate high probabilities. For example, according to the Figure 3a, among all pixels in the testing set predicted by the NN model with probability 0.80, more than 80% actually belong to the VC class. However, the overall calibration performance is good, as can be seen in Figure 3b. This figure presents the distribution of predicted probabilities, showing that for the VC class, most probabilities are concentrated at low (less than 0.2) and high (greater than 0.8) values. 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.

- L181-182: The authors state that all cases here used to validate the model were challenging because they were composed of ash, ice and SO2. I'm wondering if the case studies included in the training data were also similarly complex (i.e. ice, SO2 dominated)? Or were some of the case studies included in the training ash-rich eruption clouds?

Thank you. The case studies included in the balanced dataset including a comprehensive compilation of eruptive events, covering a wide range of volcanic cloud conditions. These events include SO_2 plumes, SO_2 dominated plumes containing minor ash components, and volcanic clouds composed exclusively of ash and SO_2. The dataset also includes challenging cases where the volcanic clouds were composed of ash, SO_2, and ice, sometimes occurring under clear-sky conditions surrounding the volcanic cloud, and in other cases in the presence of meteorological clouds surrounding or mixing with the volcanic cloud.

We have added more information in section 2.3. "Dataset" about the volcanic cloud conditions included in the balanced dataset. In addition, Table S1 was relocated to Appendix A to facilitate

faster consultation. This table provides a detailed description of the SEVIRI images employed to generate the dataset and in this new version (Table A1) we added a brief comment describing the composition of the volcanic cloud during each eruptive event.

The dataset was generated from 1.259 SEVIRI images covering 49 eruptive Etna events between December 2020 and February 2022 (see Figure 2.1). Overall, the dataset represents a comprehensive compilation of eruptive events, covering a wide range of volcanic cloud conditions. These events include SO2 plumes, SO2 dominated plumes containing minor ash components, and volcanic clouds composed exclusively of ash and SO2. The dataset also includes highly challenging cases in which the volcanic clouds were composed of ash, SO2, and ice, sometimes occurring under clear-sky conditions surrounding the volcanic cloud, and in other cases in the presence of meteorological clouds surrounding or mixing with the volcanic cloud. Table A1 provides a detailed description of the SEVIRI images employed to generate the dataset, including the number of images, the time range for each event, and a brief comment describing the composition of the volcanic cloud during each eruptive event.

- L190: It might be more informative to include a real image (alongside the schematic referred to in Fig. 2.6), to see the impact of the non-local mean filter i.e. before/after the NL mean filter is applied. The plume tracking images (Figure 4) are great.

Thank you. We have included a real image before and after the application of the NL mean filter in Figure 2f.

- L210: “values are based on typical wind speeds of 10-16 m/s” - At what height? I assume there should be a height dependence for this calculation as well.

Thank you. We have added information regarding the altitude range associated with the typical wind speeds observed during the validation events. In these cases, volcanic clouds at altitudes of 10-12 km.

The radius of the circular region increases by approximately 9-15 km (3-5 pixels) for each new image (t_{i+1}), depending on the wind speed for that day. These values are based on the SEVIRI spatial resolution and on typical wind speeds of 10-16 m/s for volcanic clouds at altitudes of 10-12 km, corresponding to the range of values observed for the validation events considered in this study. This parameter can be adjusted if required.

- L239: Evaluation metrics - I know these are standard metrics but it would be useful to have these mathematically defined in the paper, with attention to exactly how they were implemented for this study and the datasets used. E.g. were they before or after the non-local mean spatial filter? Were they only calculated for pixels falling inside the plume tracking region?

Done. Thank you. We have added a new subsection named Evaluation metrics, and described more how the metrics were used in our study.

- L262: “volcanic cloud evolution is quite similar” - Do you mean the volcanic cloud composition is quite similar? The transport and dispersion is markedly different as you point out in the next sentence.

Thank you. This sentence has been corrected.

For this event, the composition of the volcanic cloud is quite similar to the previous case.

- L270: I suggest being consistent in reporting the metrics e.g. here I would include the precision alongside the F1-score and Recall metrics to be consistent with the previous Event description. Accuracy doesn’t need to be reported here as it’s pretty much the same for all events as you state earlier.

Done. Thank you.

- L297: “4.4 Metrics analyse” -> “Metrics analysis”

Done. Thank you.

- L333-334: False positive rates - How is the false positive rate defined and calculated here?

Thank you. The false positive metric, together with all other evaluation metrics, is defined in Section 3.5 Evaluation Metrics of the revised manuscript.

- L340: “consistent with expectations” - Describe what these expectations are.

Done. Thank you.

Furthermore, as illustrated in Figure 8g-8i, the number of false negatives progressively increases over time, as expected under the hypothesis that the dilution of the volcanic cloud leads to an increase in false negatives predictions. Consequently, lower recall values are obtained. This behaviour can be observed comparing Figures 8a–8c and 8g–8i, where an increase in the number of false negatives corresponds to a decrease in the recall metrics.

- L350: “where traditional methods failed to detect volcanic clouds” - I would be a bit more specific here. Perhaps say “where traditional methods (i.e. a negative 11-12 BTD) cannot be used to detect volcanic clouds”. I think the authors need to be clear and specific here as some

of the papers cited e.g. Prata et al. (2020) deliberately used a positive BTD signature to detect volcanic clouds.

Done. Thank you.

These findings contrast with previous studies (Gupta et al., 2022; Prata et al., 2020; Rose et al., 1995; Taylor et al., 2023), where traditional methods (i.e. a negative BTD[11-12]) cannot be used to detect volcanic clouds in the presence of ice or water droplets.

- L354: VPR retrievals - I understand how this is done using the VPR method (e.g. Guerrieri et al., 2023). However, for the neural network, how are these retrievals applied to ash, ice and SO2 individually if the neural network algorithm simply classifies pixels as VC? I'm not sure I follow how this analysis was conducted.

Thank you. Following Comment #4 from Reviewer #1, as well as the present comment, a new subsection entitled *3.7. Retrieval Comparison* was included to provide a clearer explanation of how the mass loading comparison was performed.

In summary, Radiative Transfer Model (RTM) computations were conducted to obtain BTD[10.8–12.0] thresholds for discriminating between ash and ice. The procedure adopted to apply the VPR algorithm separately to each component was the same as that presented in Guerrieri et al. (2023). See Figure 6 presented below.

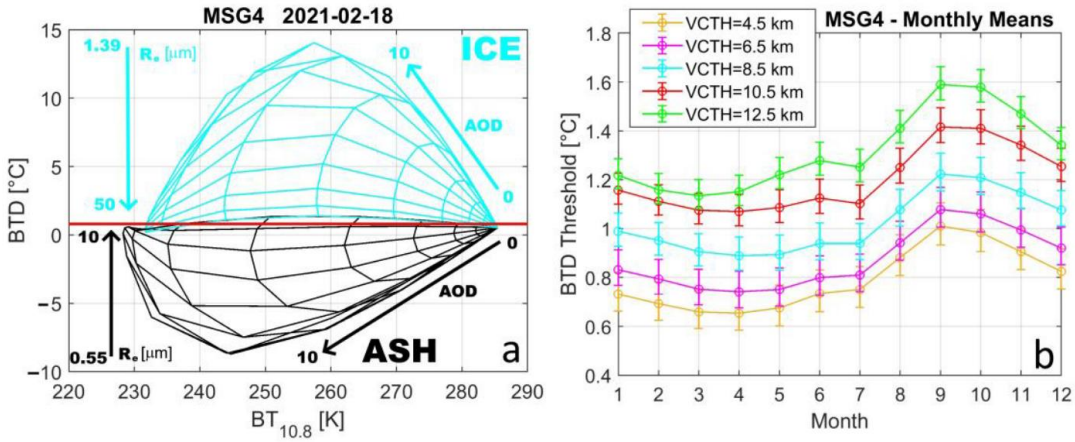


Figure 6. (**a**) Inverted arches curves of ice (cyan lines) and ash (black lines) obtained for MSG4-SEVIRI 18 February 2021 Etna LF using MODTRAN 5.3 RTM and varying $AOD_{0.55}$ and R_e. The BTD threshold (BTD = 0.8) is obtained with a horizontal linear fit (red line) of the points with maximum R_e. (**b**) Theoretical MSG4-SEVIRI BTD thresholds monthly variation as a function of VCTH of Etna and the VZA (identified by error bars).

3.7. Retrieval comparison

An additional analysis was conducted to estimate the volcanic cloud mass loading from both the manually observed data and the NN predicted data. The mass loading was retrieved using the Volcanic Plume Removal (VPR) algorithm (Pugnaghi et al., 2013, 2016), where the pixels classified by the NN model as VC class were used as input to the retrieval procedure.

Since the NN model output generates a general plume mask containing ash, ice, and SO2 without discriminating among these components, the procedure adopted to apply the VPR algorithm separately to each component was the same as that presented in Guerrieri et al. (2023). In this procedure, Radiative Transfer Model (RTM) computations were performed to derive a BTD[10.8-12.0] threshold for discriminating between ash and ice. Accordingly, a positive BTD[10.8-12.0] threshold was used to discriminate ash from ice within the NN predicted plume mask, whereas the SO2 retrieval was applied to the entire plume mask. The thresholds used for each event were 1.1514 for Event 1, 1.1218 for Event 2, and 1.2504 for Event 3. For further details regarding the calculation of the BTD[10.8-12.0] thresholds, see Sections 2.2 and 2.3 in Guerrieri et al. (2023).

The retrievals derived from the NN predicted plume masks were subsequently compared with those obtained from the manually generated plume masks (using the same BTD[10.8-12.0] thresholds).

- L378: “Figure 6 presents … SHAP values” - I think you mean Figure 10? It would also benefit the reader if the authors could add a brief explanation on how to interpret the SHAP figure. E.g. What do negative vs. positive SHAP values mean? How are “Feature Values” interpreted? e.g. is feature value normalised from 0-1? Some initial text on this would then make the physical interpretation of the channels easier to follow (e.g. L393-396).

Thank you. The figure number has been corrected. In addition, more details useful for the interpretation of the SHAP figure were added in Section 3.6.

The discussion section presents a beeswarm plot illustrating feature relevance for the NN model based on the testing set. The beeswarm plot displays the features along the y axis, ordered according to their influence on the model prediction, from the highest to the lowest. The impact of each feature on the prediction is represented along the x axis through the SHAP values, while each dot corresponds to a pixel from the testing dataset. The color of the dots indicates whether the corresponding pixel exhibits a high or low normalized value for the respective feature.

For instance, in Figure 10, the BTD[10.8–12.0] feature shows that low values, corresponding to negative BTD[10.8–12.0] values and represented in blue tones, are associated with positive SHAP values, indicating a positive contribution to the VC class prediction. In contrast, high BTD[10.8–12.0] values, corresponding to positive BTD[10.8–12.0] values, are associated with negative SHAP values, indicating a negative contribution to the VC class prediction and therefore favouring the NVC class.

- L390-392: Commentary around the BTD[10.8-12.0] - Again, I would rephrase this to say something like: the result is expected given that a negative BTD[10.8-12.0] is most sensitive to semi-transparent ash-rich volcanic clouds and the volcanic clouds considered in the present study were either opaque and/or ice/SO2-rich. I would also add the BTD[10.8-8.7] combination is used for semi-transparent ice clouds e.g. see Wang et al. (2011) (https://doi.org/10.1175/JAMC-D-11-067.1) and it appears as an important feature here because of the presence of ice in some of these volcanic clouds.

Done. Thank you.

Following in the feature relevance ranking, BTD[10.8-9.7] and BTD[10.8-8.7] appear in fourth and fifth place, respectively. The BTD[10.8–8.7] feature is commonly used for the detection of semi-transparent ice clouds (Wang et al., 2011) and appears as an important feature here because of the presence of ice in some of the volcanic clouds.

The BTD[10.8-12.0], widely recognized as the standard method for volcanic cloud detection ranks tenth, being surpassed even by the 13.3 μm channel. This result is expected, since negative BTD[10.8–12.0] values are more sensitive to semi-transparent ash-rich volcanic clouds, whereas the volcanic clouds considered in the present study were predominantly opaque and rich in ice and SO2.

- L395: Presence of SO2 in the 7.3 μm band - I agree this is probably the main reason but I would add that cold opaque clouds in general will be picked up by the 7.3 μm band. So regardless of the SO2 signal, if the plume is opaque and cold the 7.3 μm would be a good discriminator from the clear land/sea surface (you actually can see this in your Figure 11).

Done. Thank you.

The high relevance attributed to the 7.3 μm channel is likely related to the strong SO2 absorption occurring at this wavelength (Pavolonis et al., 2020), as well as to the cold and opaque characteristics of some volcanic clouds, which are effectively detected by the 7.3 μm channel. In addition, this channel provides good discrimination between volcanic clouds and clear land or sea surfaces (see Figure 11c).

This behaviour is expected because, in the presence of SO2, the 7.3 μm channel is affected by absorption. Similarly, cold or optically thick volcanic clouds produce lower brightness temperature values in this channel.

- L410-412: “BTD values around 0.0. This is expected, given that the presence of ice in volcanic clouds can compromise the performance of the BTD method.” - While I agree that ice in volcanic clouds can impact the ‘ash signal’, a BTD of 0 K between the 11-12 μm channels indicates an opaque cloud, meaning that the compositional information cannot be inferred. In other words, a BTD=0 K could be a dense ash cloud or ice cloud or liquid water cloud - we just don’t know until it becomes semi-transparent.

Thank you. This paragraph has been revised to provide a more specific and clearer explanation.

The limitation of the BTD[10.8–12.0] is evident in Figure 11m, where the volcanic cloud exhibits BTD[10.8-12.0] values close to or higher than 0.0. As previously discussed, this behaviour is expected due to the optically thick characteristics of the volcanic cloud, together with the presence of ice, both of which can reduce the effectiveness of this method for volcanic cloud detection.

- L418: If it is straightforward (within scope of the study), it would be interesting to understand (quantify) how the plume tracking improved the NN model's performance.

Thank you. The performance metrics obtained using the raw NN plume mask, together with those derived from the BTD method, were added to Table 5 for comparison.

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

- L428-429: "middle and final parts of the image sequences." -> I suggest adding further context to this statement e.g. "middle and final parts of the image sequences for the three Mt Etna eruption case studies on 22 Feb 2021, 4 March 2021 and 4 August 2024 that were analysed here" or similar.

Done. Thank you.

This limitation becomes evident when analysing the false positive rate, which increases during the middle and final parts of the image sequences for the three analysed Mt Etna eruption case studies of 22 Feb 2021, 4 March 2021 and 4 August 2024.

Figures

- Figure 1: I wonder if it would help adding coastlines on these figures to help the reader see where the land masses are in the Ash RGB image?

Done. Thank you.

- Figure 2: Should the panels be labelled (a), (b), (c) etc? Instead of 1, 2, 3.. That way in the text you would refer to "Figure 2(a)" which is more typical of the AMT journal style.

Done. Thank you.

- Figure 3: The figure caption is incomplete. Can you describe both Fig. 3(a) and (b)?

Done. Thank you.

- Figure 5: How is the purple enclosed circle calculated and how is it related to the plume tracking radius model described earlier?

The purple enclosed polygon was included just for visualization purposes, following the same approach adopted by the Volcanic Ash Advisory Centers (VAACs) for reporting volcanic clouds and by the VOLcanic Cloud Analysis Toolkit (VOLCAT) described in Pavolonis et al. (2018). However, to improve clarity and avoid potential misunderstandings, it has been removed from the figures in the revised manuscript.

Pavolonis, M. J., Sieglaff, J., and Cintineo, J.: Automated Detection of Explosive Volcanic Eruptions Using Satellite-Derived Cloud Vertical Growth Rates, Earth and Space Science, 5, 903–928, https://doi.org/10.1029/2018EA000410, 2018.

- Figures 5, 6, 7: I suggest moving the "(a)", "(b)" panels to the top right outside of the image axis bounds to improve readability. I didn't notice where these labels were at first.

Done. Thank you.

Supplementary material

The animations are really useful to have alongside the paper. Thank you for providing them.

Thank you.

References

Check the DOI links. I noticed for Guerrieri et al. (2023) for example, this link didn't work: https://doi.org/10.3390/RS15082055/S1

Done. Thank you.