GPS Solutions (2024) 28:116
https://doi.org/10.1007/s10291-024-01656-0

ORIGINAL ARTICLE

Volcanic clouds detection applying machine learning techniques to GNSS radio occultations

Mohammed Hammouti[1] · Christian Natale Gencarelli[1] · Simone Sterlacchini[1] · Riccardo Biondi[2]

Received: 17 October 2023 / Accepted: 8 April 2024

Abstract

Volcanic clouds detection is a challenge especially when meteorological clouds are present in the same area. Several algorithms have been developed to detect and monitor volcanic clouds by using satellite instruments based on different remote sensing techniques. This work aims at classifying volcanic clouds based on atmospheric profiles retrieved by the GNSS (Global Navigation Satellite Systems) radio occultation technique. We collocated the radio occultations with the volcanic cloud detection from AIRS (Atmospheric InfraRed Sounder) and IASI (Infrared Atmospheric Sounding Interferometer) for 11 big eruptions happening in the period 2008–2015 resulting in about 15000 profiles. We created an archive with the collocations and a corresponding number of profiles in "non-volcanic" environment in the same area and on the same period of the year. A support vector machine algorithm was applied to the archive in order to classify the clouds and to distinguish the volcanic clouds from the other types. The model performances are promising: the GNSS radio occultations are able to distinguish the volcanic clouds with an accuracy higher than 80% when the eruption occurs at high latitudes. The performances of the model are affected by the number of collocations used for the training. Nowadays, the number of radio occultations is higher than in the period considered in this research, making this work a pioneering study for a future operational product.

Keywords Volcanic clouds · Remote sensing · GNSS · Radio occultation · Machine learning

Introduction

Explosive volcanic eruptions present several serious hazards to society, including impacts to health (Baxter et al. 1999; Forbes et al. 2003; Horwell 2007, 2015) and safety, to life and economic assets from proximal threats (e.g., volcanic ash fall, pyroclastic density currents, lava, and toxic gases), and potential longer term deleterious effects on weather and climate (e.g., global cooling from large scale eruptions) (Robock 2000, 2013). Explosive eruptions are known for emitting large amounts of gases and aerosols, which can reach high altitudes, i.e., stratospheric layer (Robock 2000), and can last for weeks or even longer, such as Mt. Kelut eruption in 2014 (Zhu et al. 2020), and so monitoring them during and after eruption events is crucial (e.g., for aviation safety). They can emit different types of aerosols and gases into the atmosphere. The most abundant gases typically consist of water vapor, carbon dioxide (CO_2), and sulfur dioxide (SO_2), and the latter injected into the stratosphere forms fine sulfate aerosols with long residence time producing a dominant radiative effect (Robock 2000). Satellite remote sensing techniques play a key role for tracking and monitoring volcanic clouds (VCs), as they can cover large geographic areas. Satellite sensors based on Ultraviolet (UV) and Infrared (IR) technologies can provide accurate information about the dispersing volcanic aerosols and gases emissions in upper troposphere and lower stratosphere layers, but cannot provide accurate height information. Instead, active remote sensing techniques, such as CALIPSO lidar, can provide accurate height information, but have poor temporal and spatial coverage (Carn et al. 2009; Prata 2009). However, fundamental parameters of VCs, such as precise cloud top altitudes are challenging to be detected using ground based, in situ and satellite remote sensing techniques (Biondi et al. 2017). The space-based Global Navigation Satellite Systems (GNSS) Radio Occultation (RO) atmospheric remote sensing is a limb sounding satellite technique, which enables

✉ Christian Natale Gencarelli
christiannatale.gencarelli@cnr.it

[1] Institute of Environmental Geology and Geoengineering (IGAG), Italian National Research Council (CNR), Milan, Italy

[2] CIMA Research Foundation, Savona, Italy

Published online: 29 April 2024

measurement of atmospheric density structure, such as temperature, pressure, and specific humidity, in any meteorological condition, and in remote geographic areas with high vertical resolution, accuracy and precision (Kursinski et al. 1997; Yen et al 2010; Yu et al. 2014). The availability of GNSS RO data (since 2001) has been widely used for studying various atmospheric applications, and all of these technique advantages allowed it to be powerful and appreciated by the scientific community (Wickert et al. 2009; Yen et al 2010; Yu et al. 2014). For example, applications relevant to this work have used GNSS RO profiles collocated with VC maps for high vertical resolution detection and monitoring of VCs altitude (Biondi et al. 2017; Cigala et al. 2019; Tournigand et al. 2020a, 2020b).

The primary goal of this paper is to develop an automated machine learning algorithm able to discriminate the presence/absence of VCs starting from GNSS RO profiles. This algorithm is based on the support vector machine (SVM) classifier (Cortes and Vapnik 1995), a kernel-based machine learning model for classification and regression analysis. Thanks to its good theoretical foundations and excellent generalization performance, the SVM has been applied in numerous scenarios across diverse fields of science, particularly when dealing with small- to medium-sized datasets (Cervantes et al. 2020; Boateng et al. 2020). The SVM has become one of the most commonly used classification methods in recent years and has been shown by many researchers to be superior to other supervised learning methods, especially for solving practical binary classification problems (Cervantes et al. 2015, 2020; Boateng et al. 2020; Sun et al. 2005; Liang et al. 2017; Raheja et al. 2016; Bhowmik et al. 2009).

In this study, volcanic eruptions events are selected from the database created by Tournigand et al. (2020a) which includes the most significant volcanic eruption events that occurred from 2006 to 2018 and characterized by a Volcanic Explosivity Index (VEI) equal to 4 or larger. The paper is organized as follows. We first report the GNSS RO technique, the initial dataset at the base of the model as well as the analyses implemented on it to prepare the data for the SVM algorithm training; then, the results of the analysis are presented in Section "Results and discussions", showing the best runs of the model and the final model setup.

Materials and methods

This section provides a comprehensive overview of the materials and methods employed in this paper. Initially, the GNSS RO technique is presented, followed by a detailed description of the volcanic cloud and atmospheric background datasets, including information on their preprocessing steps and associated uncertainties. Lastly, the support vector machine algorithm is introduced, accompanied by a thorough explanation of the 17 experiments conducted.

GNSS remote sensing: GNSS RO technique

The GNSS RO (Kursinski et al. 1997) is a technique allowing to profile the atmospheric parameters by using the signal transmitted by a GNSS satellite and analyzed by a receiver on board of a Low Earth Orbit (LEO) satellite. The radio signal is refracted and bent in the atmosphere by the vertical density gradient, thus information about the vertical structure of the troposphere and stratosphere can be obtained. The horizontal resolution of the RO varies from about 50 km in the troposphere to 300 km in the stratosphere (Kursinski et al. 1997), while the vertical resolution varies from 100 m in the troposphere to 500 m in the stratosphere (Zeng et al. 2019).

Datasets

In this study, from each RO profile, we considered the bending angle (BA) and the temperature (T) parameters and calculated their respective anomalies (BA_{anom} and T_{anom}) as described in the following section. The BA is the most directly observable parameter in RO and contains information on the atmospheric vertical structure due to pressure, temperature, and water vapor (Biondi et al. 2011). In the lower troposphere the BA is mostly affected by the water vapor content, while in the upper troposphere and lower stratosphere (UTLS), the water vapor content decreases, and the temperature contribution prevails (Biondi et al. 2011, 2012, 2015, 2017).

As demonstrated by Biondi et al. (2017), the BA_{anom} and T_{anom}, calculated using the anomaly technique, have been demonstrated to be more effective than the BA and T parameters in detecting VC tops and their impacts on the thermal structure (Biondi, 2017; Cigala, 2019)..

In order to develop classification algorithms capable of discriminating between the presence or absence of VC using GNSS RO data, two datasets were created and used in this work: Base_dataset, and FVC_dataset (fresh volcanic cloud).

The Base_dataset consists of two classes of GNSS RO profiles:

- RO-VC, Volcanic Cloud: GNSS ROs that belong to the eruptive period. These data are selected from the multi-sensor satellite-based archive collecting all the ROs collocated with the largest SO_2 VCs since 2006 (Tournigand et al. 2020a). In particular, the volcanic eruption for the events of Okmok, Kasatochi, Sarychev, Eyjafjallajökull, Grímsvötn, Tolbachik, Nabro, Merapi, Kelut, Puyehue-Cordón Caulle (PCC), Calbuco have been analyzed, as reported in Table 1. In Fig. 1 is shown the location of the

Table 1 Eruption events data, number of RO (total of RO-VC and RO-AB) and analyzed period for Base and FVC datasets

Eruption event	Main eruptive event date	Hemis	n° RO base dataset	Period of analysis base dataset	n° RO FVC dataset	Period of analysis FVC dataset
Okmok	12/07/2008	North	894	07/12/2008–08/06/2008	432	12/07/2008–17/07/2008
Kasatochi	07/08/2008	North	10284	08/07/2008–08/29/2008	772	08/08/2008–11/08/2008
Sarychev	14/06/2009	North	4668	06/11/2009–07/16/2009	1238	11/06/2009–20/06/2009
Eyjafjallajökull	20/03/2010	North	1150	05/05/2010–05/21/2010	466	05/05/2010–13/05/2010
Grímsvötn	21/05/2011	North	1934	05/22/2011–06/18/2011	468	22/05/2011–25/05/2011
Tolbachik	27/11/2012	North	438	11/27/2012–12/03/2012	118	27/11/2012–29/11/2012
Nabro	12/06/2011	Equat	1864	05/31/2011–06/25/2011	312	12/06/2011–15/06/2011
Merapi	04/11/2010	Equat	428	10/26/2010–11/11/2010	232	05/11/2010–08/11/2010
Kelut	13/02/2014	Equat	60	02/17/2014–02/18/2014	60	17/02/2014–18/02/2014
PCC	04/06/2011	South	1148	06/07/2011–06/18/2011	352	07/06/2011–08/06/2011
Calbuco	22/04/2015	South	6930	04/24/2015–05/24/2015	844	24/04/2015–27/04/2015

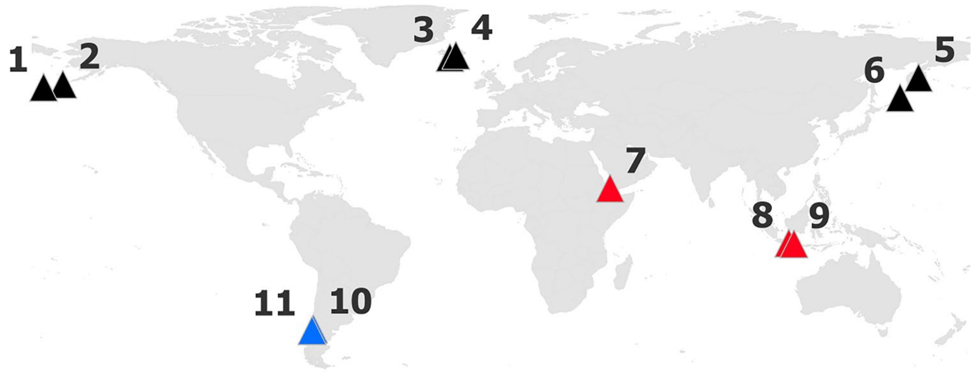

Fig. 1 Location of analyzed volcanoes: 1 Okmok, 2 Kasatochi, 3 Eyjafjallajökull, 4 Grímsvötn, 5 Tolbachik, 6 Sarychev, 7 Nabro, 8 Merapi, 9 Kelut, 10 PCC, and 11 Calbuco. The volcanoes of the northern hemisphere are shown with the black symbols, red for Equatorial area, and blue for the southern hemisphere, according with Table 1

analyzed volcanoes. As a background for the SO_2, estimations from Atmospheric InfraRed Sounder (AIRS) and Infrared Atmospheric Sounding Interferometer (IASI) data were used.

- RO-AB, Atmospheric Background: GNSS ROs that belong to the non-eruptive period. These data represent a background for ROs that belong to the non-eruptive period and are selected from the Wegener Center for Climate and Global Change (WEGC) archive (Angerer et al. 2017) with the procedure described in following section, point 2.

The number of profiles in the RO-VC datasets depends on the data availability in the Tournigand et al. (2020a). Overall, around 14900 profiles were extracted over a period of about 200 days, from the eruption to the following 35 days. The same number of RO-AB profiles were extracted to have a balance between the two classes of GNSS RO profiles during model training.

Instead, the FVC_dataset represents a subset of Base_dataset containing only the data relating to the first days of the eruption. Therefore, reducing the period between the main eruption and the RO acquisition date, it is possible to focus on the "fresh cloud" in order to better analyze its initial phase. In this case the number of ROs is obviously significantly reduced (Table 1) With approximately 2700 profiles over a period of about 40 days, from the eruption to the following 9 days.

Data pre-processing

For the creation of Base_dataset and FVC_dataset, for each volcano eruption the RO data were processed following this procedure:

1. Selecting all the RO profiles from the multi-sensor satellite-based archive (Tournigand et al. 2020a) that belong to the considered volcano (RO-VC).
2. Selecting the same number of RO profiles in non-eruptive period (RO-AB) from the WEGC archive (Angerer et al. 2017) to be used as a reference background. Particularly, for each RO-VC profile we chose the nearest one out of all the profiles within a radius of 0.5 degrees and a time range of 10 days from the RO-VC event date in a year different from the eruption event.

3. For all the selected RO-VC and RO-AB profiles, both BA and T parameters have been extracted, and their relative anomaly profiles have been calculated as follows (Biondi et al. 2011, 2012, 2017; Cigala et al. 2019):

   a. Calculation of the BA reference climatology (BA_{clim}) and T reference climatology (T_{clim}) in the same area of RO: the reference climatology is calculated by selecting BA and T profiles of all ROs collected from 2007 to 2017 and located within the same area of RO, with a radius of 2.5 degree of latitude and longitude. Then, averaging all on a monthly basis.

   b. Applying the anomaly technique to calculate the BA anomaly (BA_{anom}) as:

$$BA_{anom} = \frac{BA - BA_{clim}}{BA_{clim}} * 100 \tag{1}$$

   c. where BA is the bending angle profile into the VC, and BA_{clim} is the bending angle climatology in the same area.

   d. Similarly, the temperature anomaly (T_{anom}) has been evaluated as:

$$T_{anom} = T - T_{clim} \tag{2}$$

where T is the temperature profile, and T_{clim} is the temperature climatology.

The BA_{anom} is computed as a percentage because the absolute value of BA is really small, while the T_{anom} is computed in absolute value because it has an intrinsic importance in the atmospheric vertical structure (Biondi et al. 2017).

As a last step of data preparation, the filling values have been cleaned, the data above the altitude of 40 km have been removed (not of interest for our analysis), and the profiles have been rescaled in order to improve the model training performance. In particular the BA and T profiles have been rescaled in the range [0, 1] as they are positive values, while the BA_{anom} and T_{anom} profiles have been rescaled in the range [− 1, 1] as they are both positive and negative values. Consequently, as result of data preprocessing, for RO-VC and RO-AB four parameters have been extracted respectively: BA, T, BA_{anom}, and T_{anom}. The described procedure has been implemented by an ad hoc MATLAB algorithm.

Data uncertainties

As described in previous subsection, for the creation of Base_dataset and FVC_dataset three different instruments are combined in order to detect VCs from eruption events: GNSS RO, AIRS, and IASI. The temporal and spatial collocation between GNSS RO and AIRS or IASI represents one of the main uncertainties in this work, i.e. RO data are collocated with AIRS or IASI at ± 0.2° spatially and ± 12 h temporally (Tournigand et al. 2020a). Moreover, there is an uncertainty related to the VC detection from AIRS and IASI instruments, depending on the injected amount of aerosols erupted, and the unknown altitude and thickness of the cloud (Tournigand et al. 2020a).

Support vector machine (SVM) algorithm

The support vector machine (SVM) is a set of supervised learning methods used for common tasks in data mining, pattern recognition and machine learning (e.g., classification, regression, and outliers' detection). Especially in recent years SVM has proven to be one of the best "out of the box" classifiers, with applications in several fields of science and in real-world problems (Cervantes et al. 2015, 2020; Sun et al. 2005; Liang et al. 2017; Raheja et al. 2016; Bhowmik et al. 2009). However, the classification accuracy can be improved by increasing sample numbers (James et al. 2013; Sordo et al. 2005). In this study, the SVM is particularly a suitable algorithm for the limited size and complex nature of the dataset used. This is because it is effective in managing nonlinear relationships by employing different kernel functions (in fact the SVM is a kernel-based machine learning model) and excels in high-dimensional spaces, facilitating the identification of the hyperplane that optimally separates classes by maximizing the margin. Such characteristics enable the SVM algorithm to generalize effectively to test data while mitigating overfitting. The models created using the SVM algorithm are able to classify the RO profiles into profiles associated and not associated with VC, following this procedure: 1) selection of training and test dataset, 2) model creation based on the training dataset, 3) classification of test dataset using the produced model, 4) model performance evaluation. The best kernel for studying individual problems is to use a-priori information. Since the a-priori information is not available in this specific study, the choice of kernel is based on the characteristics of the data (Horn et al., 2018; Cervantes et al. 2020). The most accurate solutions for our binary classification problem are the 3rd-degree polynomial (poly3) and the Radial Basis Function (RBF) kernels (Nanda et al., 2018; Kasnavi et al. 2018). In the initial stages of the analysis, we also experimented with other popular kernels, such as linear and 2nd-degree polynomial kernels. However, the results were not as satisfactory as those obtained with the poly3 and RBF. Hence, we limited our focus to these two kernels to prevent the article from becoming overly complex and difficult to read.

The kernel that returns the best results on the studied datasets has been subjected to hyperparameters optimization using k-fold cross-validation technique over the training dataset to further improve the classification accuracy.

Hyperparameters optimization is used to find the best parameters that are not directly learnt within the model, and in conjunction with the k-fold cross-validation technique in order to control problems, such as reducing overfitting. In this work the parameter tuning was used applying a grid search strategy and doing fivefold cross-validation for several possible specified values of a model parameter and then choosing the parameters value with the lowest cross-validation average error.

Two different metrics have been used to evaluate the model performance in order to calculate a value for the correctly classified samples (accuracy, acc, as defined in Chicco and Jurman (2020) Eq. 1) and one that allows to balance also the false negatives (F1 score, as defined in Chicco and Jurman (2020), Eq. 4). Both assume values between 0 and 1, with 1 best value for classification. To have also an estimation of the degree of overfitting the Training Test Accuracy Rate (TTAR) has been defined as the ratio of the acc on training and on the test. Values of TTAR close to 1 indicate the absence of overfitting. All the SVM analyses and the metrics calculation have been implemented in Python 3.6.

Experiments performed

A total of 17 experiments were performed, varying the data used to train and test the model, following the indications given in Table 2. The first 8 experiments (named Okm, Kas, Sar, Eyj, Gri, Nab, PCC, Cal) refer to the analysis on single eruptions, where the profiles of each single event are considered to train the model and to validate it. In this way a specific model for each single eruption is created. In the Okm-Kas experiment the eruptions of Okmok and Kasatochi, very close in space and time (about 500 km and 27 days between the main eruption of Okmok and main eruption of Kasatochi), were considered in the same dataset to create a model capable of representing the 2 events together with a simple merging of their relative datasets. The same dataset merging operation was performed for experiments North, Equat, and South, to train models based on latitudinal events selection, while in all experiment all the events were considered. For these 13 experiments (Okm, Kas, Sar, Eyj, Gri, Nab, PCC, Cal, Okm-Kas, North, Equat, South, and All) a random 80%/20% split was used for the training/test set while simultaneously ensuring an approximately balanced distribution of both target classes (RO-VC and RO-AB profiles) in order to ensure a correct training of the model.

Additional 4 experiments were performed using some events for the training phase and different events for the test phase, with the aim of creating models that can be used for detecting other VC eruptions for which there is no data yet:

- Test1—Training on Okmok, Kasatochi and Sarychev events and testing on Eyjafjallajökull, Grímsvötn and Tolbachik events data;
- Test2—Training on Okmok and Kasatochi events and testing on Sarychev events data;
- Test3—Training on Okmok, Kasatochi and Sarychev events and testing on Nabro event data;

Table 2 Experiments description

Experiment name	Description	Type of experiment
Okm	Okmok event data	Single event
Kas	Kasatochi event data	Single event
Sar	Sarychev event data	Single event
Eyj	Eyjafjallajökull event data	Single event
Gri	Grímsvötn event data	Single event
Nab	Nabro event data	Single event
PCC	PCC event data	Single event
Cal	Calbuco event data	Single event
Okm-Kas	Okmok and Kasatochi events data	Set of 2 close events
North	Okmok, Kasatochi, Sarychev, Eyjafjallajökull, Grímsvötn and Tolbachik events data	Cluster of events
Equat	Nabro, Merapi and Kelut events data	Cluster of events
South	PCC and Calbuco events data	Cluster of events
All	All volcanic events data	Cluster of events
Test1	Training on Okmok, Kasatochi and Sarychev events and testing on Eyjafjallajökull, Grímsvötn and Tolbachik events data	Test on cluster of events
Test2	Training on Okmok and Kasatochi events and testing on Sarychev events data	Test on cluster of events
Test3	Training on Okmok, Kasatochi and Sarychev events and testing on Nabro event data	Test on cluster of events
Test4	Training on Okmok, Kasatochi and Sarychev events and testing on PCC event data	Test on cluster of events

- Test4—Training on Okmok, Kasatochi and Sarychev events and testing on PCC event data;

In these cases, the training/test ratio is not 80/20 but adjusted to the data availability: 82/18 for Test1, 71/29 for Test2, 89/11 for Test3, and 93/7 for Test4. Considering that the robustness of the model depends on the length of the training dataset, only experiments with at least about 750–800 ROs have been considered (see Table 1).

During the initial stage of analysis, we chose to train the models without a validation set (only from experiments not subjected to the cross-validation technique), intending to generally evaluate the performance and accuracy of various experiments involving different combinations of the kernel functions with all profile parameters. Instead, in a second stage, the validation set was taken into account for all experiments subjected to the cross-validation technique. Specifically, for each experiment, we employed a "Stratified 5-folds Cross Validation" strategy over the training dataset to ensure that the frequencies of the two target classes (RO-VC and RO-AB profiles) were approximately preserved in each training and validation fold.

Results and discussions

In this work the RO associated with VC data generated by the 11 largest eruptions of this century have been processed and organized into 2 groups of datasets: Base_dataset (for all data associated with eruptions) and FVC_dataset (for data related to the first days of the eruption). A total of 29800 ROs (RO-VC and RO-AB) were selected for Base_dataset while 5300 of these were selected for FVC_dataset. Each dataset reports information for 4 parameters: the bending angle (BA), the temperature (T), and their respective anomalies (BA_{anom} and T_{anom}). These datasets were used to train models (one model for each parameter, separately) based on SVM algorithms to classify if they are collocated or not with VC, representing a first attempt in detecting the VC starting from RO data.

Considering that the performance of the SVM is related to the kernel used for the classification, preliminary tests have been conducted to find the best kernel for the studied datasets. Tests were performed with poly3 (Table S1), and RBF (Table 3). The values of acc and F1 score obtained with the last 2 kernels are similar (RBF results are slightly better than poly3) but the TTAR values are higher considering poly3, as can be seen comparing Table 3 with Table S3 (a graphical comparison of the two tables is shown in Figure S2). This means that using the RBF kernel the overfit of the model is limited, thus the RBF proves to be the most appropriate kernel for the objectives of this study. Therefore, all the values reported in the paper refer to the SVM algorithm with RBF kernel, while some examples with poly3 are reported for completeness as supplementary material (Table S1 and Table S2).

In Table 3 are shown the models performance on the test sets for the anomalies of BA and T for Base_dataset. Analyzing initially the first 8 experiments (Okm, Kas, Sar, Eyj, Gri, Nab, PCC, Cal) the values of acc and F1 score are greater than 0.60. The best performances are obtained for Eyj, Gri, and Okm experiments, while the worst for

Table 3 Model results (without cross-validation) for base datasets using SVM RBF kernel

Experiment name	BA anom acc	BA anom F1 score	BA anom TTAR	T anom acc	T anom F1 score	T anom TTAR
Okm	0.82	0.80	1.07	0.74	0.74	1.16
Kas	0.68	0.71	1.10	0.69	0.71	1.08
Sar	0.61	0.62	1.17	0.65	0.65	1.09
Eyj	0.85	0.85	1.06	0.84	0.84	1.08
Gri	0.78	0.79	1.08	0.82	0.83	1.04
Nab	0.76	0.76	1.14	0.74	0.75	1.14
PCC	0.74	0.77	1.08	0.77	0.80	1.05
Cal	0.74	0.76	1.07	0.75	0.75	1.05
Okm-Kas	0.69	0.71	1.08	0.70	0.72	1.05
North	0.63	0.64	1.09	0.66	0.68	1.04
Equat	0.72	0.72	1.15	0.76	0.77	1.08
South	0.67	0.69	1.13	0.73	0.73	1.05
All	0.61	0.62	1.11	0.63	0.64	1.06
Test1	0.46	0.31	1.55	0.46	0.34	1.52
Test2	0.56	0.56	1.33	0.57	0.56	1.30
Test3	0.54	0.50	1.31	0.47	0.43	1.48
Test4	0.57	0.49	1.25	0.57	0.57	1.23

Sar. In the cases with good acc (range from 0.74 to 0.84 for Eyj, Gri, and Okm) of the model the number of FP and FN is low, therefore no wrong overestimation/underestimation of VC is shown in these models (Powers 2020).

The values of acc and F1 score reported in Table 3 demonstrate that the developed models can correctly classify the ROs associated with the VCs for individual events, on par with those from other previous similar studies (Torrisi et al 2022; Cervantes et al. 2020).

Good model performances have been found both on the anomalies of T and BA, while the acc and the F1 score decreases when considering the absolute values of BA and T. Specifically the acc decreases by about 11% by comparing the performances of BA and BA_{anom} and about 12% by comparing T and T_{anom} (see Table 3 and Tables S1, S2, S3). Also in this study, the anomaly technique (Biondi et al. 2017) proves effective on both BA and T profiles.

The same performances of Sar (worst case for the accuracy of the models) were obtained by training the model on all volcanoes (All experiment), while grouping events by latitude returns more satisfactory results (acc between 0.72 and 0.76 for Equat experiment). In general, training the model considering multiple events together leads to worse results. This is evident by comparing the separate Okm and Kas experiments with the Okm-Kas experiment, which shows a lowering of the accuracy in the Okm-Kas experiment, especially when BA and T are considered in absolute value.

As additional proof, experiments in which the SVM model has been trained on past events and tested on events that have not yet occurred were conducted. These are the experiments labeled Test1 and Test4 in Table 2. In Test 1 the north latitude events prior to 2010 were used for model training (Okm, Kas and Sar) while those after 2010 for the test (Eyj, Grim and Tol). The same models were tested on 2011 events (Test 3 on Nab, Test 4 on PCC), while in Test 2 the north latitude events prior to 2009 were used for model training (Okm and Kas) and that of 2009 for the test (Sar). Unfortunately, in terms of acc and F1 score, the results are not satisfying in this last set of experiments: the models constructed in this way overestimate the cases of false positives (the number of real positives that are wrongly predicted as negative), which means that the models tend to underestimate the VCs on the test. Another evidence is the strong overfitting (high TTAR values), which demonstrates that further investigations are needed to carry out similar analysis.

The low performances of Tests 1–2 demonstrate that the algorithm must be customized at regional scale due to different factors:

- The reference climatology used to compute the anomaly is different according to the latitude—moving towards higher (lower) latitudes, the tropopause height decreases (increases) and this affects the computation of the anomaly at different layers;
- A higher frequency of convection in the area decreases the tropopause temperature;
- Each volcano is usually characterized by a specific type of eruption (e.g. mainly SO2, mainly ash, water vapor rich clouds, mixed clouds, …) affecting in different way the atmospheric structure in terms of density and radiative effect;
- In some cases there can be a combination of different clouds (e.g. 2 eruptions or a volcanic eruption during convective activity).

In order to improve the model performance with a further increase of classification acc for each experiment, an optimization of the SVM hyperparameters C and γ was performed for the BA and T anomaly profiles. The parameters have been optimized by cross-validated grid-search over a parameter grid as described in section "Support Vector Machine (SVM) algorithm". As expected, an improvement in model acc was found in each experiment (Table 4) up to 4% in BA_{anom} in the Gri experiment (acc from 0.78 without Cross Validation to 0.81 with Cross Validation) and 10% in T_{anom} in the Kas experiment (acc from 0.69 without Cross Validation to 0.76 with Cross Validation). However, the experiments Test1—Test4 have been excluded from the parameter optimization training processes considering the low performances obtained.

The structure and properties of a VC over time can be affected by meteorological and atmospheric factors as it disperses in the atmosphere. Considering a FVC that belongs to the first days of volcanic eruption, and possibly near to the volcano geographic area, may increase the reliability and robustness of the classification accuracy results. Consequently, for each volcanic eruption event it has been selected from the archive Tournigand et al. (2020a) only the profiles that belong to the first days of the eruption event, and thus obtaining the FVC_dataset (a subset of Base_dataset) as explained in previous section and Table 2.

Analogous to the Base_dataset analysis, the same study was also repeated on FVC_dataset training SVM models on BA_{anom} and T_{anom} profiles with RBF kernel. However, considering that the FVC_dataset refers only to the first few days following the eruption, some events have a small number of profiles, and so they have not been considered. Only the following experiments have been analyzed: Kas, Sar, Cal, Okm-Kas, North, South, All. The produced models with default parameters based on the BA_{anom} profiles showed acc values between 0.70 and 0.76 in the Cal, Kas, North, Okm-Kas and Sar, low values form the All and South experiments (between 0.63 and 0.66). Similar values were also found for T_{anom} profiles, with acc ranging from 0.70 to 0.81 for the Cal, Kas, North, Okm-Kas and Sar experiments, and lows values in the All and South cases (0.67 and 0.65 respectively), as shown in Table 5.

Lastly, hyperparameters optimization along with k-fold cross-validation technique have been applied for the experiments related to FVC_datasets. The results are shown in Table 6. As the Base_dataset analysis case, also here the created models from BA_{anom} and T_{anom} profiles showed an increase in the acc and F1 score values.

The performance of the proposed machine learning algorithm is comparable to other algorithms based on machine learning techniques (Torrisi et al. 2022), but it also represents the first attempt to create a model working at global scale and not for case studies (Corradini et al. 2010; Torrisi et al. 2022; Piontek et al., 2021; Corradini et al. 2021; Romeo et al. 2023). The algorithm's performance can be improved, e.g., by setting other specific thresholds for the BA_{anom} and T_{anom} profiles, but this will be the subject of future investigations, as it is not the objective of this study. Additionally, this algorithm has the advantage of not depending on other parameters and models as it happens for the Brightness Temperature Difference algorithm (Corradini et al. 2010, 2021; Prata and Lynch 2019; Romeo et al. 2023)

Table 4 Model results with cross-validation for Base datasets using SVM RBF kernel

Experiment name	BA anom acc	BA anom F1 score	BA anom TTAR	T anom acc	T anom F1 score	T anom TTAR
Okm	0.82	0.80	1.09	0.77	0.76	1.19
Kas	0.71	0.72	1.25	0.76	0.77	1.19
Sar	0.62	0.63	1.22	0.66	0.67	1.09
Eyj	0.88	0.88	1.10	0.86	0.86	1.12
Gri	0.81	0.82	1.10	0.84	0.85	1.10
Nab	0.77	0.77	1.21	0.75	0.76	1.20
PCC	0.78	0.79	1.10	0.81	0.82	1.16
Cal	0.75	0.76	1.08	0.77	0.77	1.18
Okm-Kas	0.71	0.73	1.18	0.75	0.76	1.17
North	0.66	0.67	1.17	0.70	0.70	1.19
Equat	0.75	0.75	1.17	0.78	0.79	1.18
South	0.70	0.71	1.20	0.75	0.75	1.20
All	0.63	0.63	1.20	0.67	0.67	1.19

Table 5 Model results (without cross-validation) for FVC datasets using SVM RBF kernel

Experiment name	BA anom acc	BA anom F1 score	BA anom TTAR	T anom acc	T anom F1 score	T anom TTAR
Kas	0.76	0.76	1.14	0.81	0.79	1.07
Sar	0.72	0.74	1.15	0.71	0.73	1.12
Cal	0.71	0.69	1.20	0.79	0.77	1.05
Okm-Kas	0.73	0.75	1.17	0.78	0.80	1.04
North	0.70	0.72	1.11	0.70	0.71	1.09
South	0.63	0.65	1.29	0.65	0.67	1.27
All	0.66	0.68	1.13	0.67	0.68	1.09

Table 6 Model results with cross-validation for FVC datasets using SVM RBF kernel

Experiment name	BA anom acc	BA anom F1 score	BA anom TTAR	T anom acc	T anom F1 score	T anom TTAR
Kas	0.77	0.77	1.15	0.83	0.82	1.08
Sar	0.73	0.75	1.20	0.76	0.77	1.12
Cal	0.72	0.68	1.14	0.83	0.80	1.12
Okm-Kas	0.74	0.75	1.12	0.80	0.82	1.10
North	0.72	0.73	1.20	0.73	0.73	1.15
South	0.65	0.64	1.17	0.68	0.65	1.17
All	0.68	0.69	1.19	0.70	0.71	1.18

for which in most cases, it is also necessary the supervision of an operator to discriminate the components of VC.

Moreover, this work uses an "uncommon" data source (GNSS RO) for this type of studies supporting the necessity of using "potential new satellites and instruments dedicated to monitoring volcanic ash plumes and eruptions due to the urgent need to gather information on the vertical structure of evolving VC" (Zehner et al., 2010) and using a reliable detection system not dependent on the meteorological conditions necessary to have a weather independent warning capacity as suggested by Tupper et al. (2004).

Conclusions

The work shows a first classification study to classify the VC starting from GNSS RO data. Based on the validation of the models, the SVM algorithm with the RBF kernel function showed good performance in most case studies, especially working with BA and *T* anomaly profiles on single eruption events. Model acc decreases if more events are considered to train the models, which suggests that further investigations are needed to carry out analysis on event clusters. It is interesting to note that the resulting accuracy from BA ($\mathrm{BA}_{\mathrm{anom}}$) and *T* ($T_{\mathrm{anom}}$) showed different values but with small differences. This could be explained by the algorithm's better performance in detecting the anomaly signatures at higher altitudes, as they are more distinguishable, such as in the UTLS, where VCs typically reach, with minimal or negligible water vapor content.

This study shows that the GNSS RO profiles are able to distinguish the VC from other atmospheric conditions. The use of the anomaly provides a performance improvement up to 15% (acc from 0.63 to 0.72 in the case of PCC experiment) depending on the volcano and this is due to the fact that the anomaly highlights the density variation of the atmospheric layers in which the VC lies. However, it is not possible to state at the moment when the $\mathrm{BA}_{\mathrm{anom}}$ accuracy is better than the T_{anom} accuracy or vice-versa. This will be the topic of future investigations.

The RO-AB were selected in different years of the volcanic eruption in a random environment to build a realistic and robust reference background. The results could show a further relevant improvement if the RO-AB were selected in a clear sky environment when the atmospheric profile approximately follows the climatology, but this can alter the model accuracy robustness in presence of dense meteorological clouds. The presence of the VCs in environments prone to convective activity can be the reason of the different performances of the algorithm for single eruptions. Eyj, Gri, and Okm clouds were just at high latitudes (Figure S1) where the convection is rare, while Sarychev (lowest performance) is in the area where the typhoons become extratropical cyclones (Biondi et al. 2015) and really strong convection can happen affecting the BA and *T* profiles in a similar manner.

The algorithm must be customized due to the atmospheric vertical profile structure changing with longitude and (mostly) latitude, so the model can provide the best performance when applied regionally. This is the main reason why the performances of the Test1, Test2, Test3 and Test4 are low and the performances increase when working on latitudinal bands.

The use of SVM algorithm based on RBF kernel with optimized hyperparameters C and γ for the anomaly profiles shows an improvement in classification acc accuracy for most of the performed experiments detecting the respective VCs with a good accuracy. Hyperparameter optimization has also contributed to improvements in terms of acc working on event clusters (e.g. Okm-kas, South, and Equat experiments). Moreover, the experiments based on the FVC_dataset showed a similar results trend to those based on the Base_dataset. Therefore, even a limited amount of data in the first few days following the event is enough to have good performance.

The number of ROs acquired in the period 2008–2015 is really small compared to the actual number of ROs available nowadays (Ho et al. 2022), and this provides a good prospective to increase the performance of the model in the future and to potentially include the GNSS RO profiles into already existing early warning system for monitoring volcanic clouds.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10291-024-01656-0.

Acknowledgements the authors thank the Wegener Center (WEGC) for making the GNSS-RO data OPSv5.6 available, accessible via https://doi.org/https://doi.org/10.25364/WEGC/OPS5.6:2019.1

Author contributions M.H.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization. C.N.G.: methodology, validation, formal analysis, investigation, writing—original draft preparation, writing—review and editing, visualization, supervision. S.S.: writing—review and editing, supervision, funding acquisition. R.B.: conceptualization, methodology, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, supervision, project administration, funding acquisition.

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement. M. H., C. N. G., S. S.: no founding. R. B.: the work was accomplished in the frame of the VESUVIO project (Volcanic clouds dEtection and monitoring for Studying the erUption impact on climate and aVIatiOn) funded by the Supporting Talent in ReSearch (STARS) grant at Università degli Studi di Padova, Italy.

Availability of data and materials The GNSS RO data collocated with Volcanic Clouds (RO-VC) were downloaded from https://doi.org/https://doi.org/10.5880/fidgeo.2020.016 (Tournigand et al., 2020a).

The GNSS RO data that belong to the non-eruptive period (RO-AB) were downloaded from https://doi.org/https://doi.org/10.25364/WEGC/OPS5.6:2019.1 (Angerer et al. 2017).

Declarations

Conflict of interests The authors declare no competing interests.

Consent for publication Not applicable.

References

Achirul Nanda M, Boro Seminar K, Nandika D, Maddu A (2018) A comparison study of kernel functions in the support vector machine and its application for termite detection. Information 9(1):5

Angerer B, Ladstädter F, Scherllin-Pirscher B, Schwärz M, Steiner AK, Foelsche U, Kirchengast G (2017) Quality aspects of the Wegener Center multi-satellite GPS radio occultation record OPSv5.6. Atmos Measurement Techn 10(12):4845–4863

Baxter PJ et al (1999) Cristobalite in volcanic ash of the soufriere hills volcano, montserrat: hazards implications. Science 283(5405):1142–1145

Bhowmik TK, Ghanty P, Roy A, Parui SK (2009) SVM-based hierarchical architectures for handwritten Bangla character recognition. Int J Doc Anal Recognit (IJDAR) 12(2):97–108

Biondi R, Neubert T, Syndergaard S, Nielsen J (2011) Measurements of the upper troposphere and lower stratosphere during tropical cyclones using the GPS radio occultation technique. Adv Space Res 47(2):348–355

Biondi R, Randel WJ, Ho SP, Neubert T, Syndergaard S (2012) Thermal structure of intense convective clouds derived from GPS radio occultations. Atmos Chem Phys 12(12):5309–5318

Biondi R, Steiner AK, Kirchengast G, Rieckh T (2015) Characterization of thermal structure and conditions for overshooting of tropical and extratropical cyclones with GPS radio occultation. Atmos Chem Phys 15(9):5181–5193

Biondi R, Steiner AK, Kirchengast G, Brenot H, Rieckh T (2017) Supporting the detection and monitoring of volcanic clouds: a promising new application of Global Navigation Satellite System radio occultation. Adv Space Res 60(12):2707–2722

Boateng EY, Joseph O, Daniel AA (2020) Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. J Data Anal Inf Proc 8(4):341–357

Carn SA, Krueger AJ, Krotkov NA, Yang K, Evans K (2009) Tracking volcanic sulfur dioxide clouds for aviation hazard mitigation. Nat Hazards 51:325–343

Cervantes J, Lamont FG, López-Chau A, Mazahua LR, Ruíz JS (2015) Data selection based on decision tree for SVM classification on large data sets. Appl Soft Comput 37:787–798

Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A (2020) A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 408:189–215

Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21(1):1–13

Cigala V, Biondi R, Prata AJ, Steiner AK, Kirchengast G, Brenot H (2019) GNSS radio occultation advances the monitoring of volcanic clouds: the case of the 2008 Kasatochi eruption. Remote Sens 11(19):2199

Corradini S, Merucci L, Prata AJ, Piscini A (2010) Volcanic ash and SO2 in the 2008 Kasatochi eruption: Retrievals comparison from different IR satellite sensors. J Geophys Res Atmos. https://doi.org/10.1029/2009JD013634

Corradini S, Guerrieri L, Brenot H, Clarisse L, Merucci L, Pardini F, Prata AJ, Realmuto VJ, Stelitano D, Theys N (2021) Tropospheric volcanic SO2 mass and flux retrievals from satellite. The Etna december 2018 eruption. Remote Sens 13(11):2225

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learn 20:273–297

Forbes L, Jarvis D, Potts J, Baxter PJ (2003) Volcanic ash and respiratory symptoms in children on the island of Montserrat British West Indies. Occup Environ Med 60:207–211. https://doi.org/10.1136/oem.60.3.207

Ho SP, Pedatella N, Foelsche U, Healy S, Weiss JP, Ullman R (2022) Using radio occultation data for atmospheric numerical weather prediction, climate sciences, and ionospheric studies and initial results from COSMIC-2, commercial RO data, and recent RO missions. Bull Am Meteor Soc 103(11):E2506–E2512

Horn D, Demircioğlu A, Bischl B, Glasmachers T, Weihs C (2018) A comparative study on large scale kernelized support vector machines. Adv Data Anal Classif 12:867–883

Horwell CJ (2007) Grain-size analysis of volcanic ash for the rapid assessment of respiratory health hazard. J Environ Monit 9:1107–1115. https://doi.org/10.1039/B710583P

Horwell C, Baxter P, Kamanyire R (2015) Health impacts of volcanic eruptions Global Volcanic Hazards and Risk. Cambridge University Press, Cambridge, pp 289–294. https://doi.org/10.1017/CBO9781316276273.015

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York

Kasnavi SA, Aminafshar M, Shariati MM, Kashan NEJ, Honarvar M (2018) The effect of kernel selection on genome wide prediction of discrete traits by Support Vector Machine. Gene Reports 11:279–282

Kursinski ER, Hajj GA, Schofield JT, Linfield RP, Hardy KR (1997) Observing Earth's atmosphere with radio occultation measurements using the global positioning system. J Geophys Res Atmos 102(D19):23429–23465

Liang X, Zhu L, Huang DS (2017) Multi-task ranking SVM for image cosegmentation. Neurocomputing 247:126–136

Piontek D, Bugliaro L, Schmidl M, Zhou DK, Voigt C (2021) The new volcanic ash satellite retrieval vacos using msg/seviri and artificial neural networks: 1. development. Remote Sens 13(16):3112

Prata AJ (2009) Satellite detection of hazardous volcanic clouds and the risk to global air traffic. Nat Hazards 51:303–324

Prata F, Lynch M (2019) Passive earth observations of volcanic clouds in the atmosphere. Atmosphere 10(4):199

Raheja JL, Mishra A, Chaudhary A (2016) Indian sign language recognition using SVM. Pattern Recognit Image Anal 26:434–441

Robock A (2000) Volcanic eruptions and climate. Rev Geophys 38(2):191–219

Robock A (2013) The latest on volcanic eruptions and climate. EOS Trans Am Geophys Union 94(35):305–306

Romeo F, Mereu L, Scollo S, Papa M, Corradini S, Merucci L, Marzano FS (2023) Volcanic cloud detection and retrieval using satellite multisensor observations. Remote Sens 15(4):888

Sordo M, Zeng Q (2005) On sample size and classification accuracy: A performance comparison. In: Biological and Medical Data Analysis: 6th International Symposium, ISBMDA 2005, Aveiro, Portugal, November 10-11, 2005. Proceedings 6. Springer Berlin Heidelberg, pp 193–201

Sun BY, Huang DS, Fang HT (2005) Lidar signal denoising using least-squares support vector machine. IEEE Signal Process Lett 12(2):101–104

Torrisi F, Amato E, Corradino C, Mangiagli S, Del Negro C (2022) Characterization of volcanic cloud components using machine learning techniques and SEVIRI infrared images. Sensors 22(20):7712

Tournigand PY et al (2020a) A multi-sensor satellite-based archive of the largest SO_2 volcanic eruptions since 2006. Earth Syst Sci Data 12(4):3139–3159

Tournigand PY et al. (2020b) The 2015 Calbuco volcanic cloud detection using GNSS radio occultation and satellite lidar. In: IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, pp 6834–6837. IEEE

Tupper A et al (2004) An evaluation of volcanic cloud detection techniques during recent significant eruptions in the western “Ring of Fire.” Remote Sens Environ. https://doi.org/10.1016/j.rse.2004.02.004

Wickert J et al (2009) GPS radio occultation: results from CHAMP, GRACE and FORMOSAT-3/COSMIC. Terr Atmos Ocean Sci 20(1):35–50

Yen NL, Fong CJ, Chu CH, Miau JJ, Liou YA, Kuo YH (2010) Global GNSS Radio Occultation Mission for Meteorology, Ionosphere & Climate. Aerospace Technologies Advancements, edited by Arif, TT, pp 241–258

Yu K, Rizos C, Burrage D, Dempster AG, Zhang K, Markgraf M (2014) An overview of GNSS remote sensing. EURASIP J Adv Signal Proc 2014:1–14

Zehner C (Ed.) (2010) Monitoring volcanic ash from space. In: Proceedings of the ESA-EUMETSAT workshop on the 14 April to 23 May 2010 eruption at the Eyjafjöll volcano, South Iceland. Frascati, Italy, pp 26–27 May 2010, ESAPublication STM-280

Zeng Z, Sokolovskiy S, Schreiner WS, Hunt D (2019) Representation of vertical atmospheric structures by radio occultation observations in the upper troposphere and lower stratosphere: comparison to high-resolution radiosonde profiles. J Atmos Oce Techn 36(4):655–670

Zhu Y, Toon OB, Jensen EJ, Bardeen CG, Mills MJ, Tolbert MA, Yu P, Woods S (2020) Persisting volcanic ash particles impact stratospheric SO2 lifetime and aerosol optical properties. Nat Commun 11(1):4526

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Mohammed Hammouti received his MSc degree in Geoinformatics Engineering from Polytechnic University of Milan in 2020. Currently, he is a research fellow at the National Research Council of Italy. His main research areas of interest are remote sensing, GIS, and machine learning applied to geosciences.

Christian Natale Gencarelli received his PhD in physics of complex systems from University of Calabria in 2013. Currently, he is a researcher at the National Research Council of Italy. Previously, he was a data analyst and IT consultant for several companies. His research activity focuses on the study of atmospheric dynamics and the application of machine learning methods to geosciences.

Simone Sterlacchini did MSc in Earth Sciences in 1992 and PhD in Engineering Geology in 1997. Since 1998, he is a researcher at the National Research Council of Italy, Institute of Environmental Geology and Geoengineering. Currently, he is scientific responsible for the Laboratory of Risk Analysis and Emergency Management, Centre of Competence of the Italian Red Cross for Emergency Planning and Decision Support. Since 2002, he is a professor at the University of Milan, teaching GIS courses.

Riccardo Biondi is a project manager at CIMA Research Foundation. He received the master degree in electronic engineering at University of Perugia (Italy) and the PhD in atmospheric physics at the Technical University of Denmark (Denmark). His research activity focuses on analysis and monitoring of extreme atmospheric events, such as volcanic eruptions, tropical cyclones, and thunderstorms, by using different remote sensing techniques.