

Article

The New Volcanic Ash Satellite Retrieval VACOS Using MSG/SEVIRI and Artificial Neural Networks: 2. Validation

Dennis Piontek ^{1,*}, Luca Bugliaro ¹, Jayanta Kar ^{2,3}, Ulrich Schumann ¹ , Franco Marengo ^{4,5}, Matthieu Plu ⁶  and Christiane Voigt ^{1,7} 

- ¹ Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, 82234 Weßling-Oberpfaffenhofen, Germany; luca.bugliaro@dlr.de (L.B.); ulrich.schumann@dlr.de (U.S.); christiane.voigt@dlr.de (C.V.)
 - ² Science Systems and Applications Inc., Hampton, VA 23666, USA; jayanta.kar@nasa.gov
 - ³ Science Directorate, NASA Langley Research Center, Hampton, VA 23666, USA
 - ⁴ Space Applications and Nowcasting, Met Office, Exeter EX1 3PB, UK; franco.marengo@metoffice.gov.uk
 - ⁵ Climate and Atmosphere Research Centre (CARE-C), Cyprus Institute, Aglantzia, Nicosia 2121, Cyprus
 - ⁶ Centre National de Recherches Météorologiques (CNRM), University of Toulouse, Météo-France, CNRS, 31057 Toulouse, France; matthieu.plu@meteo.fr
 - ⁷ Institut für Physik der Atmosphäre, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany
- * Correspondence: dennis.piontek@dlr.de

Abstract: Volcanic ash clouds can damage aircrafts during flight and, thus, have the potential to disrupt air traffic on a large scale, making their detection and monitoring necessary. The new retrieval algorithm VACOS (*Volcanic Ash Cloud properties Obtained from SEVIRI*) using the geostationary instrument MSG/SEVIRI and artificial neural networks is introduced in a companion paper. It performs pixelwise classifications and retrieves (indirectly) the mass column concentration, the cloud top height and the effective particle radius. VACOS is comprehensively validated using simulated test data, CALIOP retrievals, lidar and in situ data from aircraft campaigns of the DLR and the FAAM, as well as volcanic ash transport and dispersion multi model multi source term ensemble predictions. Specifically, emissions of the eruptions of Eyjafjallajökull (2010) and Puyehue-Cordón Caulle (2011) are considered. For ash loads larger than 0.2 g m^{-2} and a mass column concentration-based detection procedure, the different evaluations give probabilities of detection between 70% and more than 90% at false alarm rates of the order of 0.3–3%. For the simulated test data, the retrieval of the mass load has a mean absolute percentage error of ~40% or less for ash layers with an optical thickness at $10.8 \mu\text{m}$ of 0.1 (i.e., a mass load of about $0.3\text{--}0.7 \text{ g m}^{-2}$, depending on the ash type) or more, the ash cloud top height has an error of up to 10% for ash layers above 5 km, and the effective radius has an error of up to 35% for radii of $0.6\text{--}6 \mu\text{m}$. The retrieval error increases with decreasing ash cloud thickness and top height. VACOS is applicable even for overlaying meteorological clouds, for example, the mean absolute percentage error of the optical depth at $10.8 \mu\text{m}$ increases by only up to ~30%. Viewing zenith angles $>60^\circ$ increase the mean percentage error by up to ~20%. Desert surfaces are another source of error. Varying geometrical ash layer thicknesses and the occurrence of multiple layers can introduce an additional error of about 30% for the mass load and 5% for the cloud top height. For the CALIOP data, comparisons with its predecessor VADUGS (operationally used by the DWD) show that VACOS is more robust, with retrieval errors of mass load and ash cloud top height reduced by $>10\%$ and $>50\%$, respectively. Using the model data indicates an increase in detection rate in the order of 30% and more. The reliability under a wide spectrum of atmospheric conditions and volcanic ash types make VACOS a suitable tool for scientific studies and air traffic applications related to volcanic ash clouds.

Keywords: volcanic ash cloud; passive satellite remote sensing; artificial neural network; validation; Eyjafjallajökull; Puyehue-Cordón Caulle; lidar; in situ; transport and dispersion model



Citation: Piontek, D.; Bugliaro, L.; Kar, J.; Schumann, U.; Marengo, F.; Plu, M.; Voigt, C. The New Volcanic Ash Satellite Retrieval VACOS Using MSG/SEVIRI and Artificial Neural Networks: 2. Validation. *Remote Sens.* **2021**, *13*, 3128. <https://doi.org/10.3390/rs13163128>

Academic Editor: Carmine Serio

Received: 1 July 2021

Accepted: 3 August 2021

Published: 7 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A new volcanic ash retrieval using *artificial neural networks* (ANNs) and the *Spinning Enhanced Visible and Infrared Imager* (SEVIRI) aboard the *Meteosat Second Generation* (MSG) satellites is developed and presented; this algorithm is called VACOS (*Volcanic Ash Cloud Properties Obtained from SEVIRI*) and builds upon its predecessor VADUGS (*Volcanic Ash Detection Using Geostationary Satellites* [1]). The companion paper [2] describes the algorithm development: Using a comprehensive set of volcanic ash optical properties [3], surface emissivities [4–6] and atmospheric profiles of pressure, temperature, air density, concentrations of oxygen, water vapor, ozone, carbon dioxide and nitrogen dioxide, liquid and ice water clouds, mostly derived from ECMWF model reanalyses, one-dimensional radiative transfer calculations are performed with and without realistic volcanic ash clouds to create training, validation and test data sets. These three simulated data sets contain the volcanic ash cloud properties, i.e., geometrical vertical extent, mass volume concentration, cloud top height, the brightness temperatures (BTs) of the infrared channels of SEVIRI and various auxiliary quantities. The ash-free simulations are validated by comparing the results of radiative transfer calculations of a specific date with the corresponding SEVIRI measurements. Using the simulated data sets, four different ANNs are trained for the pixelwise retrieval of the optical depth at $10.8\ \mu\text{m}$ due to ash ($\tau_{10.8}$), the ash cloud top height (in m, z_{top}), the effective particle radius (in μm , r_{eff}) and an overall classification in four categories (ash-free and cloud-free; only meteorological clouds; only volcanic ash clouds; both volcanic ash and meteorological clouds present). The ANN for classification returns a normalized four-dimensional vector, where each component can be roughly interpreted as the probability of the corresponding category. The four ANNs perform independently of each other, but the retrievals of z_{top} and r_{eff} receive an estimate of $\tau_{10.8}$ as an input. This approach allowed to use different training data sets and ANN settings for each retrieval.

This paper contains an analysis of the retrieval performance: A detailed validation with respect to simulated test data sets is presented (Section 2) and the sensitivity of the retrievals with respect to the volcanic ash cloud profile is given (Section 3). To demonstrate the reliability of the new algorithm and to check its performance with respect to its predecessor, various comparisons with other remote and in situ measurements (Section 4) and model calculations were made (Section 5). The individual features of the final ANNs are analyzed to make some inferences on the functioning of the algorithms (Section 6). Finally, we give a conclusion and an outlook.

2. Performance on Simulated Test Data

The development of the VACOS retrieval is described in Piontek et al. [2]. In the following, we systematically quantify the performance of the retrievals with respect to volcanic ash cloud properties, presence of meteorological clouds (defined to include liquid and ice water clouds) and geographic location. Therefore, the ANNs are applied to the test data sets A (1,252,470 samples) or B (405,556 samples) from Piontek et al. [2], depending on their training data set. The samples of the test data sets are the results of independent radiative transfer calculations and can be compared to the situation for single pixels in a SEVIRI image. The VACOS results are compared with these true values, providing references for the error of the retrievals; those might be larger in reality due to more complicated atmospheric conditions (e.g., additional aerosols such as mineral dust), cases that have not been covered by the training data set (e.g., non-homogeneous ash clouds or multiple ash layers, emitted sulfur dioxide) or slight differences between our radiative transfer calculations and the reality (e.g., due to partial cloud covers, inaccuracies due to the applied parameterizations for meteorological clouds). The error metrics *mean absolute percentage error* (MAPE), *mean percentage error* (MPE), *probability of detection* (POD), *false alarm rate* (FAR) and *accuracy* are used and described in Appendix A.

2.1. Classification

The classification ANN returns a normalized four-dimensional vector with each component interpreted as the probability of the corresponding category, see Table 1. Defining that a sample is assigned to a given category if the corresponding component is $>50\%$, the accuracy is 92% with 0.012% of the samples remaining unclassified. These values depend on the composition of the test data set with respect to the different categories, as they are retrieved with different accuracy, see Table 1. Whereas clear sky, cloudy and ash-loaded samples are correctly classified with probabilities of more than 90% each, samples with both meteorological and ash clouds are correctly classified in only 49% of the cases. About 47% of those samples were classified as ash only. Again, this might depend on the composition of the test data; samples with thick ash clouds on top of comparably thin meteorological clouds might be misclassified as ash only. Reducing the amount of these samples in the test data set would significantly modify the results in Table 1. Nevertheless, VACOS is able to detect the presence of ash in almost all cases even for this category, with only 4.1% of the ash remaining undetected. Next, only the samples containing ash and meteorological clouds are considered and separated according to the location of the meteorological clouds as *above* or *below* the ash layer, where we define that *above* denotes that no meteorological clouds are below the volcanic ash cloud bottom and *below* that no meteorological cloud is above the ash cloud top. Samples with multiple meteorological clouds located both above and below the ash layer are not included in either class. Again, note that we do not differentiate between liquid and ice water clouds, although ice water clouds can be expected to dominate for altitudes in the upper troposphere and liquid water clouds in the lower troposphere, and although ice water clouds might damp the ash signal, e.g., in BT_{11-12} . Furthermore, the dependence on the optical depth of the meteorological clouds themselves is not investigated, although expected to be significant. Table 2 shows the classification of the two subsets. Nearly all samples are classified as ash-containing if the meteorological cloud is below, and astonishingly still $\sim 85\%$ if it is above. The amount of correctly classified samples (i.e., both ash and meteorological cloud) is $\sim 15\%$ higher for meteorological clouds below than above, whereas $\sim 15\%$ of the samples are classified as containing only meteorological clouds if those are above but less than 1% if they are below. This represents the well known fact that an optically thick meteorological cloud can effectively hide a below-cloud volcanic ash layer from the satellite observation. More than 48% of the samples are classified as containing only ash, independently of the position of the meteorological clouds. Motivated by the fact that the identification of cases with both ash and meteorological clouds is not very reliable, but that ash is detected in the majority of the situations, a binary ash flag P_{ash} (i.e., ash or no ash) is introduced by adding the probabilities of the two categories without ash (*clear*, *clouds*) and the two with ash (*ash*, *both*), respectively. Now, if the resulting probability for ash is above 80%, we assume that ash is present, otherwise not; the threshold is motivated in Section 2.3. The binary ash flag will be used in the rest of the section. It has an accuracy of 99.5%, a POD of 98.6% and a FAR of 0.008% for the simulated data.

Table 1. Results of the classification ANN with respect to the simulated test data in percent; four categories are differentiated: ash-free and cloud-free (*clear*), only meteorological clouds (*clouds*), only volcanic ash (*ash*), both volcanic ash and meteorological clouds (*both*); the true value is given in the left column, the corresponding number of samples and how they are classified is given in the other columns.

Truth	Samples	Retrieval/%			
		Clear	Clouds	Ash	Both
clear	560,713	99.7	0.3	<0.1	<0.1
clouds	287,740	5.6	94.3	<0.1	<0.1
ash	279,395	<0.1	<0.1	94.6	5.2
both	124,622	<0.1	4.1	46.8	49.2

Table 2. Results of the classification ANN with respect to the simulated test data in percent; only samples with volcanic ash and meteorological clouds are considered; *above* denotes that no meteorological clouds are below the volcanic ash cloud bottom and *below* that no meteorological cloud is above the ash cloud top; the retrieval categories are the same as in Table 1.

Cloud Location	Samples	Retrieval/%			
		Clear	Clouds	Ash	Both
above	21,833	<0.1	15.1	48.8	36.1
below	81,630	<0.1	0.3	48.0	51.6

2.2. Dependence on Volcanic Ash Cloud Properties, Meteorological Clouds and Geographic Coordinates

Next we analyze the binary ash flag and the regression retrievals in detail. As metrics we calculate the accuracy for the binary ash flag, and MAPE and MPE for the retrievals of $\tau_{10.8}$, z_{top} and r_{eff} for simulated samples within certain intervals of the true values $\tau'_{10.8}$, z'_{top} or r'_{eff} (in the rest of this work, primed quantities will always denote the reference data, which might be the truth when using simulated data, or in situ measurement/retrieval/model results in the other cases). As discussed in Piontek et al. [2], $\tau'_{10.8}$ can be converted to a mass column concentration m'_{col} using the mass extinction coefficient at 10.8 μm , with a mean value of $\sim 200 \text{ m}^2 \text{ kg}^{-1}$. Thus, the investigated range $\tau'_{10.8} \in [0.01, 10]$ corresponds approximately to $m'_{\text{col}} \in [0.05, 50] \text{ g m}^{-2}$. $m'_{\text{col}} = 0.05 \text{ g m}^{-2}$ is quite low; typical mass loads are about one order of magnitude larger (Section 4). Figure 1 shows subsets with land and sea surfaces, and with and without meteorological clouds. Figure 2 shows results for ash-containing samples with meteorological clouds above and below the ash layer (as defined before). Test data set A is considered for the binary ash flag when investigating the dependence on $\tau'_{10.8}$, otherwise only the ash-loaded samples are used. For the retrieval of $\tau_{10.8}$ only the ash-loaded samples of test data set A are used, and for the retrieval of z_{top} and r_{eff} the test data set B, which also contains only ash-loaded samples [2]. Note that no prior selection is made based on whether or not ash is detected in a sample using the binary ash flag or the retrieved $\tau_{10.8}$. The sample distribution for test data set B is given in Figures 1h and 2h; generally, the distribution is similar for test data set A with differences of <10%, except for the first bin which contains also the ash-free samples when considering the accuracy with respect to $\tau'_{10.8}$.

For the binary ash flag, high accuracies of 90–100% are found for usual ash clouds ($\tau'_{10.8} > 0.1$) and in absence of volcanic ash (left-most bin in Figure 1a1). The additional presence of meteorological clouds decreases the accuracy: if they are above the ash layer the difference is of the order of 20%, whereas the influence is much smaller when they are below. This demonstrates that ash layers might often be hidden by the meteorological clouds above. The accuracy is also close to 100% for $z'_{\text{top}} > 5 \text{ km}$, but decreases with decreasing z'_{top} . The latter might be partly connected to the impact of water vapor above the ash cloud, as their column load above the ash layer increases with decreasing z'_{top} . However, as the accuracy decreases only slightly for $z'_{\text{top}} < 5 \text{ km}$ compared to higher z'_{top} in the absence of meteorological clouds, the impact of water vapor appears to be limited. The presence of meteorological clouds leads to a much worse performance, especially for $z'_{\text{top}} < 5 \text{ km}$ and if the meteorological clouds are above the ash layer, where the accuracy drops well below 50%. The dependence on r'_{eff} is generally small, except when meteorological clouds are present above, leading to a decreasing accuracy with $\sim 90\%$ for $r'_{\text{eff}} = 0.6 \mu\text{m}$, but less than 60% for $r'_{\text{eff}} = 6 \mu\text{m}$. The dependence on the surface (land/sea) is small.

For the regression ANNs, the MAPE generally decreases from roughly 100% for $\tau'_{10.8} = 0.03$ / $z'_{\text{top}} = 1 \text{ km}$ to less than 30% for $\tau'_{10.8} = 10$ / $z'_{\text{top}} = 18 \text{ km}$ with increasing $\tau'_{10.8}$ and z'_{top} , and is up to 35% with respect to r'_{eff} . In all cases, the MAPE is smallest in the absence of meteorological clouds and largest in their presence, with significantly larger errors for meteorological clouds above with respect to $\tau'_{10.8} > 0.3$ and r'_{eff} , and for meteorological clouds below with respect to z'_{top} . The MAPEs for land and sea surfaces are again rather similar.

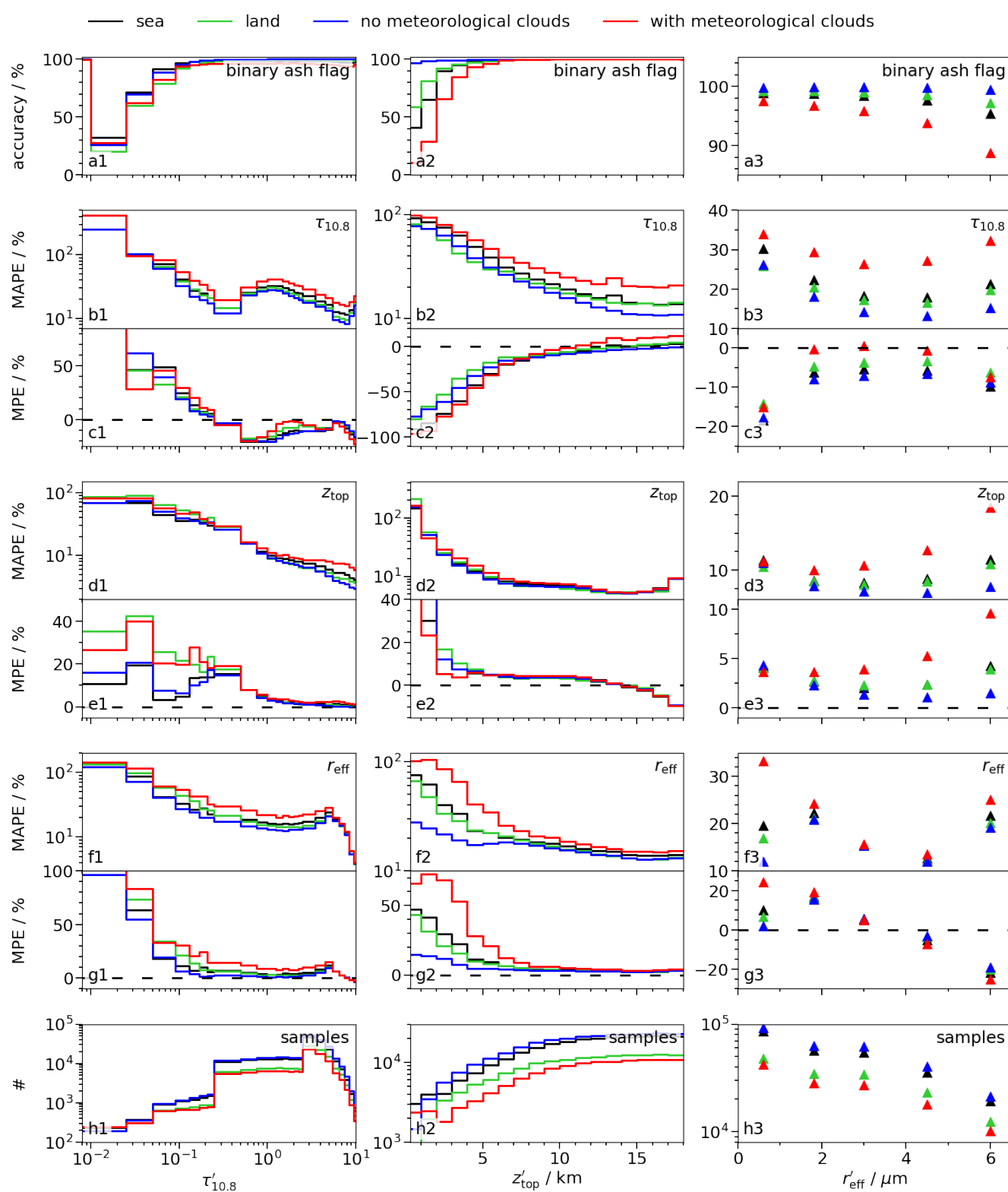


Figure 1. Estimations of the accuracy of the binary ash flag (a), and MAPE and MPE for the retrievals of $\tau_{10.8}$ (b,c), z_{top} (d,e) and r_{eff} (f,g) based on the simulated test data sets; the columns show the performance for different $\tau_{10.8}$ (1), z_{top} (2) and r_{eff} (3); for (a–c) the test data set A is used, for (d–g) test data set B [2]; for (b,c) only ash-loaded samples are considered; the sample distribution of test data set B is shown in (h); different subsets are shown, i.e., only sea (black) or land (green) surfaces, only samples with meteorological clouds (red) or without meteorological clouds (blue).

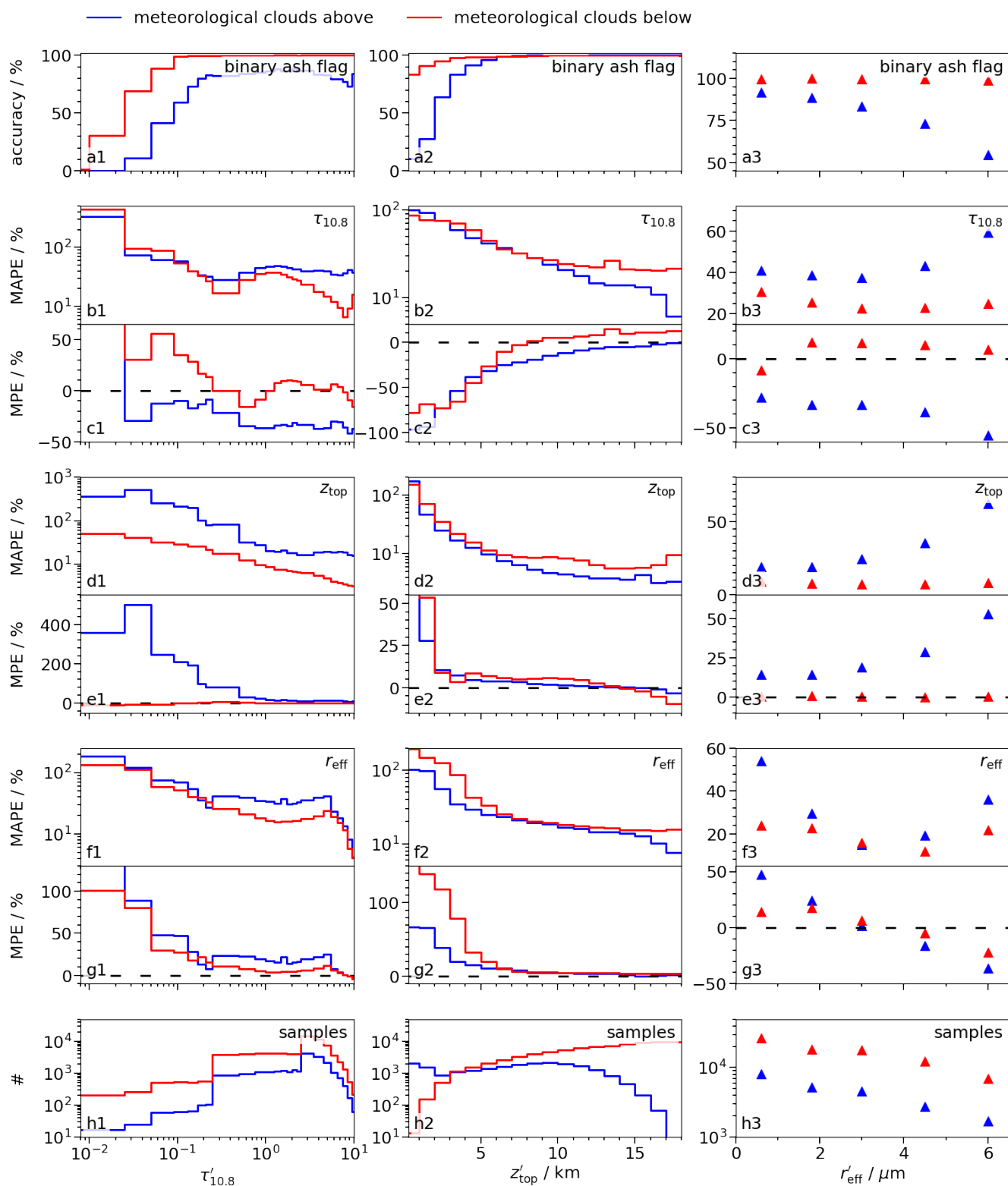


Figure 2. Similar as in Figure 1: Estimations of the accuracy of the binary ash flag (a), and MAPE and MPE for the retrievals of $\tau_{10.8}$ (b,c), z_{top} (d,e) and r_{eff} (f,g) based on the simulated test data sets; the columns show the performance for different $\tau'_{10.8}$ (1), z'_{top} (2) and r'_{eff} (3); for (a–c) the test data set A is used, for (d–g) test data set B [2]; for (b,c) only ash-loaded samples are considered; the sample distribution of test data set B is shown in (h); only samples with meteorological clouds above (blue) and below (red) the volcanic ash layer are considered.

On average, the retrieval of $\tau_{10.8}$ is biased towards high values for $\tau'_{10.8} < 0.25$ (i.e., $\text{MPE} > 0$) and towards low values otherwise ($\text{MPE} < 0$); the MPE has values of -20 to 60% for $\tau'_{10.8} > 0.025$. The MAPE has minima for $\tau'_{10.8}$ around 0.3 and 8 ; the first minimum might be explained by the increased sample weights applied during training for $\tau'_{10.8}$ at corresponding values [2]. For $z'_{\text{top}} < 10$ km, the retrieved $\tau_{10.8}$ is underestimated with the MPE being -50 to -100% for $z'_{\text{top}} < 4$ km and 0 to -50% for $4 \text{ km} < z'_{\text{top}} < 10$ km.

The retrieved z_{top} is generally overestimated, with MPEs up to 50% but below 20% for $\tau'_{10.8} > 0.5$ and below 10% for $z'_{\text{top}} > 4$ km. The bias is larger for optically thin ash clouds (e.g., $\tau'_{10.8} < 0.17$) with meteorological clouds above, whereas nearly no bias is apparent in the presence of meteorological clouds below. For $z'_{\text{top}} > 15$ km, the retrieval error of z_{top} slightly increases again; a physical reason might be that at high latitudes the tropopause is located at similar heights, such that the disappearance or inversion of the vertical temperature gradient makes the determination of z_{top} more difficult.

The retrieved r_{eff} is overestimated for all values of $\tau'_{10.8}$ and z'_{top} ; the MPE becomes $>50\%$ for $\tau'_{10.8} < 0.05$ and between 10% and more than 140% for $z'_{\text{top}} < 3$ km, strongly depending on the presence of meteorological clouds. For $r'_{\text{eff}} < 4 \mu\text{m}$, the retrieved r_{eff} is overestimated, and underestimated beyond; the MPE is generally between -20% and 20% .

Our simulated test data sets consist of samples which are calculated for specific geographical locations. The latitude/longitude coordinates are drawn randomly, but are equally distributed with respect to the SEVIRI disc [2]; thus, more samples are located around 0°N , 0°E than at larger viewing zenith angles. The georeferenced test data are used to investigate the dependence of the accuracy, POD and FAR of the binary ash flag on the geographical position in Figure 3, and the MAPE and MPE of the regression ANNs in Figure 4. In all cases, four different subsets are investigated: with meteorological clouds, without meteorological clouds, with meteorological clouds above and below (as defined before). The samples are arranged in boxes of $10^\circ \times 10^\circ$, except for the FAR in Figure 3c1,c2, which is given for $45^\circ \times 45^\circ$ boxes to accumulate enough samples given its small numerical value. The decreasing sample density towards the edges of the plots can explain the worse retrieval performances at higher viewing zenith angles. Thus, we focus on the central regions of $\pm 40^\circ$ around 0°N , 0°E .

The binary ash flag has a high accuracy and POD (both close to 100%) in the absence of clouds except for the desert regions of Northern Africa and the Arabian peninsula, where both metrics decrease by $1\text{--}2\%$, whereas the FAR rises to about 0.008% . This might be connected to the surface emissivity of quartz-rich soils that can lead to a negative $\text{BTD}_{10.8-12}$ [7]. In the presence of clouds, the accuracy and POD remain close to 100% above Africa, especially above the tropical forest and in proximity of the equator; otherwise the metrics decrease by up to 3% and 10% , respectively. The decrease is even more pronounced if meteorological clouds are above, which might be connected to the fact that the simulated samples are mostly located above tropical Africa; thus, the ANNs are in this case mainly exposed to samples of a very specific type. Meteorological clouds below are predominantly above the south east Atlantic off the coast of Africa, where an extensive marine stratus deck at low altitudes is usual [8]. The non-uniform distribution of the different sample types resembles reality, as the occurrence of meteorological clouds in the radiative transfer simulation is based on their presence in the ECMWF model [2]. To try to improve the performance of the ANNs for the physically less common cases, one could increase their amount in the training data set in the future, by specifically selecting those samples during the data set creation, or by increasing their sample weight during the training. However, this would also distort the underlying probability distributions, e.g., by artificially increasing the number of samples with meteorological clouds above the ash layer above the Atlantic, the corresponding probability would be higher for the training data set than in reality. A priori it is not clear whether this would improve the overall performance of the algorithm. Note that other retrievals exhibited similar properties as the binary ash flag, e.g., the ice cloud retrieval CiPS had the highest POD above forest for cirrus clouds with an optical thickness up to 0.5 , and at the same time a high FAR above equatorial Africa as cirrus clouds often occur in this region (i.e., above tropical

rainforests) and, thus, corresponding samples made up a significant part of the used training data set [9,10].

The retrieval of $\tau_{10.8}$ has a MAPE of mostly 20–30%, independently of meteorological cloud presence. Again there is a high MAPE of more than 60% above the Atlantic west of southern Africa if meteorological clouds are above the ash. $\tau_{10.8}$ is generally underestimated with MPEs of 0 to −20%, except if meteorological clouds are below the ash layer, which leads to MPEs of 0–15%.

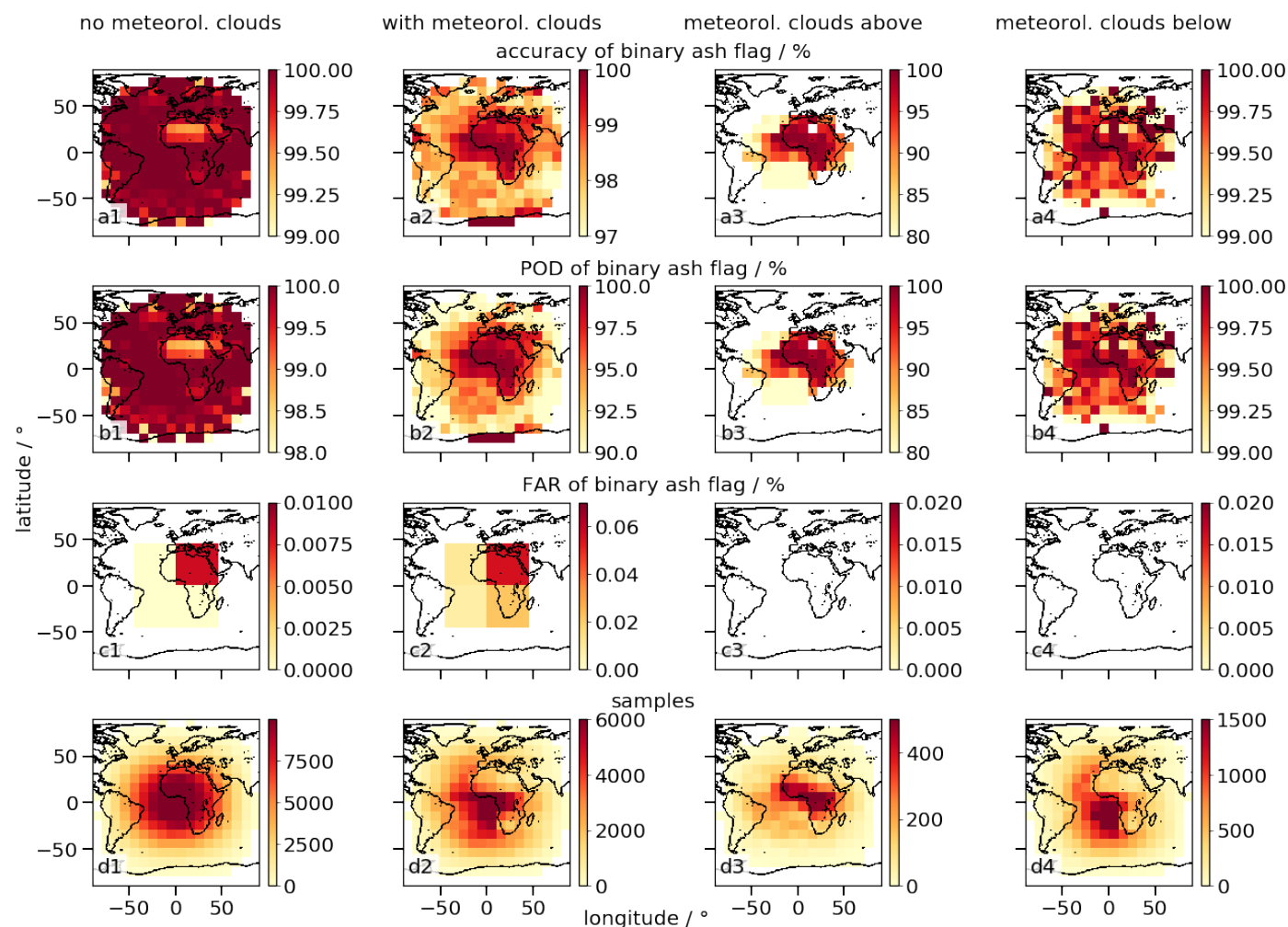


Figure 3. Estimations of the accuracy (a), the POD (b) and the FAR (c) for the binary ash flag for different geographical coordinates and subsets without meteorological clouds (1), with meteorological clouds (2), with meteorological clouds above (3) and below (4) a volcanic ash layer; the number of samples is given in (d); no metrics are shown if a grid cell contains <100 (<50,000 for FAR) samples; mind the different color scales.

The retrieval of z_{top} shows a latitudinal dependence of the MAPE: it is 5–10% at the equator, but rises towards the poles up to 20% at $\pm 50^\circ\text{N}$. Similarly, the MPE rises from $\sim 0\%$ around the equator to about 10%. A reason for the latitudinal dependence of MAPE and MPE could be that the ANN might learn mainly the vertical temperature profile at the tropics, as there the sample density in the training data is the highest. Since the temperatures are generally lower towards the poles [11], the retrieved z_{top} from the measured brightness temperatures using the tropical temperature profile is systematically too high. Theoretically, this issue might be overcome by using a training data set of samples equally distributed over the complete latitudinal range, again either by accordingly increasing the amount of samples or the weight of the given samples. Then there should be comparable focus on all the different temperature profiles. The ANN's decision on which

temperature profile to use could be based on the latitude, but also on, for instance, the time of day and day of year, all of which are used as input data for VACOS. However, another problem might be the decreasing height of the tropopause and the corresponding vertical temperature inversion at higher latitudes, which makes it harder to deduce z_{top} from the measured brightness temperatures for ash clouds above the tropopause. Differences in the surface emissivity due to ice and snow surfaces [5] might also introduce errors in the most poleward regions. The last two issues would not be solved by using a training data set evenly distributed across latitudes.

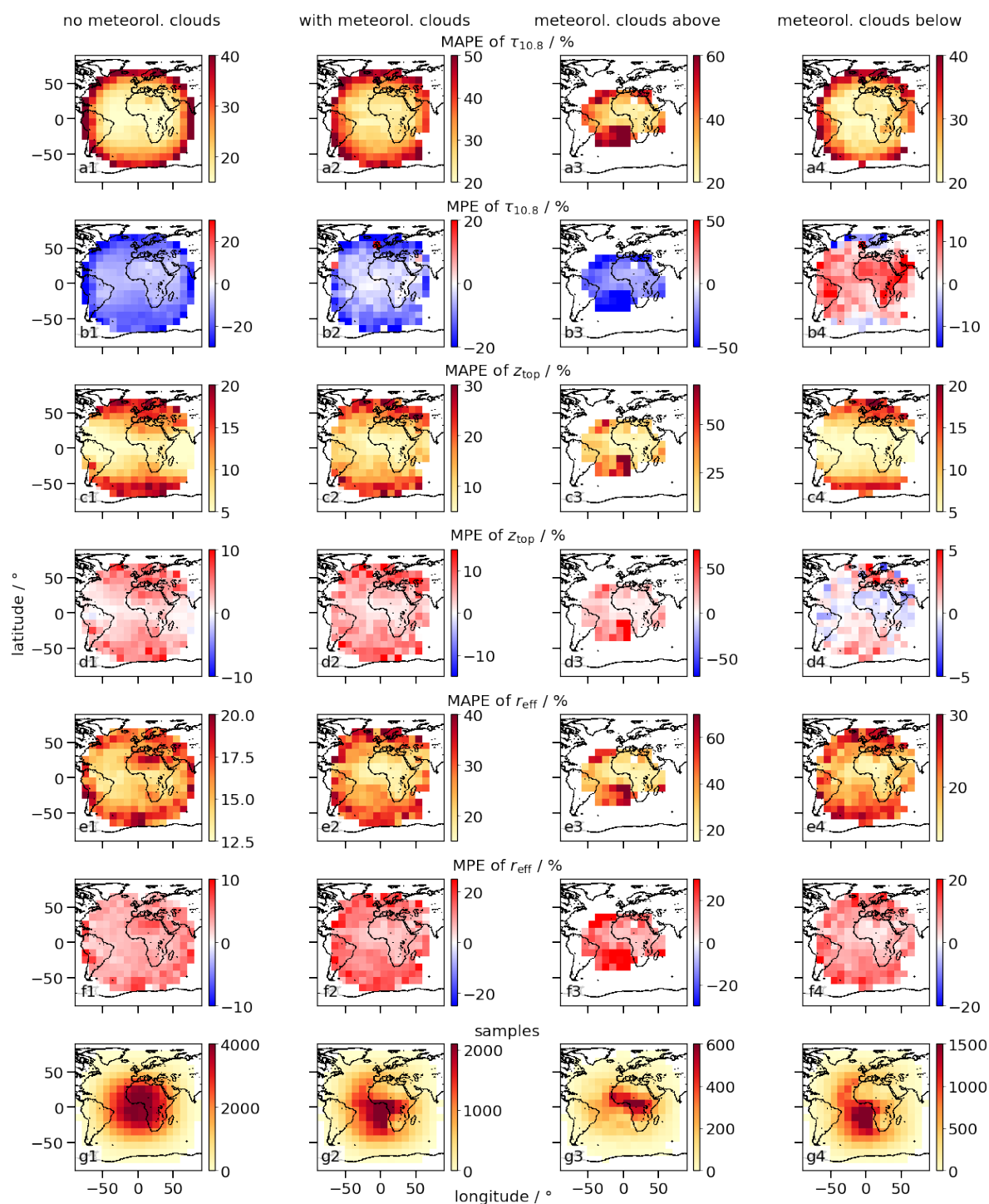


Figure 4. As in Figure 3, but given are MAPE and MPE for the retrieval for $\tau_{10.8}$ (a,b), z_{top} (c,d) and r_{eff} (e,f) and the number of samples (g) according to test data set B.

The retrieval of r_{eff} is generally overestimated, with MAPEs at the equator around 15% if no meteorological clouds are present, and 20% on average in their presence. Again, the error is higher above the Sahara in the absence of meteorological clouds.

2.3. Detection of Volcanic Ash

Volcanic ash can be detected using the following threshold rules: a sample is classified as ash-containing if one of $BTD_{12-10.8}$, $\tau_{10.8}$ or the probability for ash due to the binary ash flag P_{ash} is larger than a given threshold $BTD_{12-10.8,\text{thrs}}$, $\tau_{10.8,\text{thrs}}$ or $P_{\text{ash,thrs}}$, respectively. Notice that the first possibility is independent of VACOS. Figure 5 shows POD and FAR for all three methods for different thresholds. Generally, by increasing the threshold the POD as well as the FAR decrease. Thus, it is necessary to find a trade-off between both properties. In case of doubt a higher POD is favored with regard to the relevance of ash detection in aviation-security. As thin ash layers are harder to detect (Section 2), two subsets are considered besides the full test data set: samples with $\tau'_{10.8} < 0.2$ and $\tau'_{10.8} < 0.5$. Using $BTD_{12-10.8}$ (e.g., [7,12,13]) with $BTD_{12-10.8,\text{thrs}} = 0 \text{ K}$ gives a POD of 100% and a FAR of about 10%. The former is due to the selection of ash samples in the simulated data sets, i.e., only ash-loaded samples with $BTD_{10.8-12} < 0$ have been considered [2]. For $BTD_{12-10.8,\text{thrs}} = 1.58$, the POD is ca. 70% and the FAR is about 0.25% for the full test data set. For the subsets of lower $\tau'_{10.8}$, the POD is much lower, whereas the FAR remains high. Using $\tau_{10.8,\text{thrs}}$ leads to higher PODs compared to $BTD_{12-10.8,\text{thrs}}$ for all three sets of test data, the curves move towards the upper left corner of Figure 5. Again the PODs are smaller for lower $\tau'_{10.8}$. For example, $\tau_{10.8,\text{thrs}} = 0.04$ leads to a FAR of 1% and a POD of 98.7% (95.4%, 86.8%) for the full test data set (only $\tau'_{10.8} < 0.5$, < 0.2). Using P_{ash} leads to an even better performance in all subsets: up to roughly $P_{\text{ash,thrs}} = 0.8$, the POD decreases only slightly, whereas the FAR decreases by multiple orders of magnitude. Only for higher $P_{\text{ash,thrs}}$, the POD drops as well. For $P_{\text{ash,thrs}} = 0.8$, the FAR is 0.008% and the POD is 98.6% (95.5%, 88.2%) for the full test data set (only $\tau'_{10.8} < 0.5$, < 0.2).

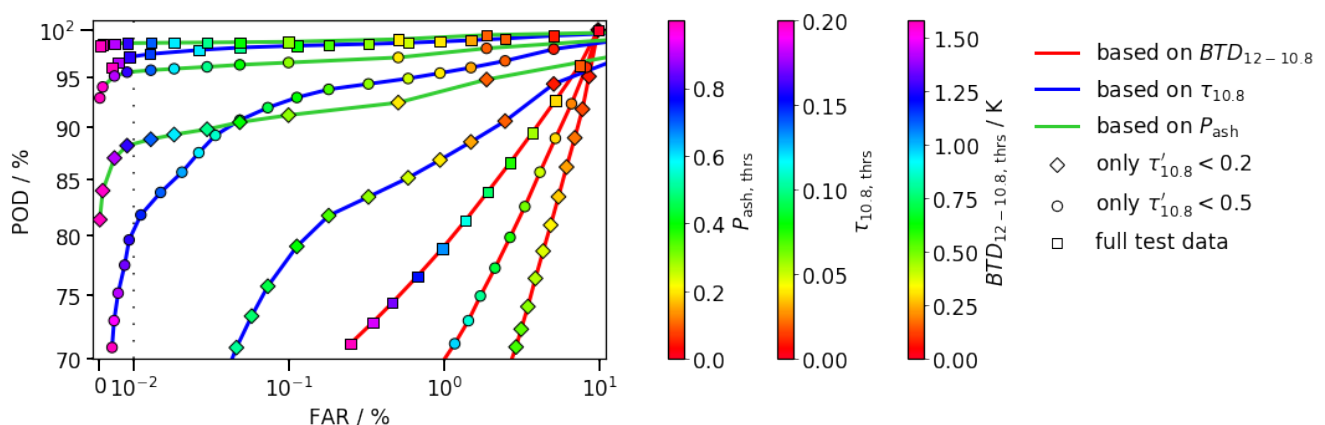


Figure 5. Estimations of the POD and the FAR for the detection of volcanic ash using $BTD_{12-10.8}$ (connected by a red line), $\tau_{10.8}$ (blue line) and the probability of the binary ash flag for ash P_{ash} (green line), and a corresponding threshold (color of marker, see colorbars); different subsets of the test data set [2] are used and encoded in the marker type, i.e., the full data set (square), only samples with $\tau'_{10.8} < 0.5$ (circle) and < 0.2 (diamond); the x-axis is linear left of the black, dotted, vertical line and logarithmic right of it.

Instead of subsets defined by the $\tau'_{10.8}$ as in Figure 5, one can also select sets according to m'_{col} : For $\tau_{10.8,\text{thrs}} = 0.04$, corresponding to 0.2 g m^{-2} when considering a mean mass extinction coefficient at $10.8 \mu\text{m}$ of $200 \text{ m}^2 \text{ kg}^{-1}$ [2], and $m'_{\text{col}} \in [0.2, 1] \text{ g m}^{-2}$, a typical regime for distal volcanic ash clouds (Section 4), the POD is ca. 93% and the MAPE for m_{col} roughly 40%, whereas $m'_{\text{col}} \in [1, 10] \text{ g m}^{-2}$ leads to a POD of 99% and a MAPE of 26%. For the ash-free samples, the FAR is about 1%.

3. Sensitivity to Volcanic Ash Cloud Profiles

For the training data set, we made the ad hoc assumption of a single, homogeneous volcanic ash layer with z_{top} up to 18 km, geometrical thicknesses up to 7.2 km and m_{col} up to 30 g m^{-2} [2]. Here we consider the sensitivity of the brightness temperatures and the ANNs on the ash cloud profile. In all cases we apply the refractive index of Eyjafjallajökull ash [14], a log-normal particle size distribution with $r_{\text{eff}} = 0.6 \mu\text{m}$, a geometric standard deviation $s = 1.5$ and a representative shape distribution of spheroids as considered in Piontek et al. [3]. A thick ash cloud is assumed with $m_{\text{col}} = 10 \text{ g m}^{-2}$, corresponding to $\tau_{10.8} = 2$ for a mass extinction coefficient at $10.8 \mu\text{m}$ of $200 \text{ m}^2 \text{ kg}^{-1}$. This order of magnitude of m_{col} can be found in close proximity of a volcano (Section 5); for lower m_{col} , the absolute impact of the investigated macrophysical properties on the brightness temperatures is assumed to be smaller.

Using ECMWF ERA5 data for 2010 and the methods described in Piontek et al. [2], a random set of 500 atmospheric and geographical conditions is chosen and used for the calculation of the brightness temperatures for each cloud setting. Only cases without meteorological but with volcanic ash clouds are investigated as meteorological clouds were already shown to influence the retrievals significantly.

3.1. Multiple Ash Layers

To quantify the variability in the brightness temperatures due to multiple layers, we compare a single layer with four different multi-layer structures, Figure 6a. All four setups keep z_{top} and m_{col} fixed (i.e., the parameters that are retrieved by the ANNs). Layers #1 and #2 also keep the mass concentration c fixed and introduce a gap in the volcanic ash cloud. Layers #3 and #4 keep the cloud bottom height fixed but increase c . The brightness temperatures of the infrared channels of SEVIRI are simulated and the differences compared to the single layer are calculated, Figure 6b. Layers #1 and #2 show that the brightness temperature increases with a decreasing height for the lower ash layer; the differences are generally positive, as more of the mass is in warmer parts of the atmosphere. For a gap of 0.5 km the differences can be $>0.5 \text{ K}$, but for a gap of 2 km the differences are even $>2 \text{ K}$, with outliers even $>3 \text{ K}$. For the more condensed structures #3 and #4, the differences are mostly between 0 K and -0.5 K and negative as more of the mass is in cooler parts of the atmosphere. Figure 6c shows the relative differences in the retrievals with respect to the true values. The median z_{top} retrieval is overestimated by ca. 3%, the retrieved r_{eff} is underestimated by about 7%. In both cases is the influence of multi-layer structures compared to single layers on the order of 1%. The median retrieved $\tau_{10.8}$ for a single layer is about 20% higher than the true value. Again structure #2 leads to the largest difference: here the median retrieved $\tau_{10.8}$ is about 10% lower than the true value. The structures #3 and #4 lead only to minor differences compared to the single layer.

3.2. Non-Homogeneous Ash Profiles

The assumption of a uniform ash layer [2] is often not fulfilled in reality, where vertical ash mass profiles do not have discontinuities and might have a clear peak [15–17]; thus, a more realistic description would be a normal distribution or the Π -sigmoid distribution (Appendix B and [18]), which has the ability to approximate uniform and normal distributions as limiting cases, Figure 7a. To quantify the influence of the cloud profile, different assumptions are compared. A uniform cloud with a typical height of 8–9 km is assumed [2]; thus, the mean height is 8.5 km and the standard deviation 0.289 km. These values are assumed for the other distributions as well. For the Π -sigmoid distribution, the variables parameterizing the lower and upper end of the cloud are varied (Π -sigmoid #1: 8.01 km and 8.99 km, #2: 8.05 km and 8.95 km, #3: 8.1 km and 8.9 km). The continuous distributions are cut off at 7.5 km and 9.5 km; within this regime they are modeled by 100 sublayers of depth 0.02 km, each having a mass volume concentration corresponding to m_{col} times the density of the normalized vertical mass distribution at the mean height of the sublayer. The brightness temperatures of the infrared channels of SEVIRI are simulated and the differences of the various profiles compared

to the uniform distribution are calculated, Figure 7b. The absolute differences are mostly <0.1 K and, therefore, smaller or of the same order as the instrumental noise [19]. Outliers are <-0.2 K. In some channels (e.g., $BT_{8.7}$ or $BT_{9.7}$) the differences for all profiles are mostly negative, indicating that a more realistic cloud profile might lead to slightly lower brightness temperatures. Applying the ANNs to the different profiles leads to negligible differences (not shown).

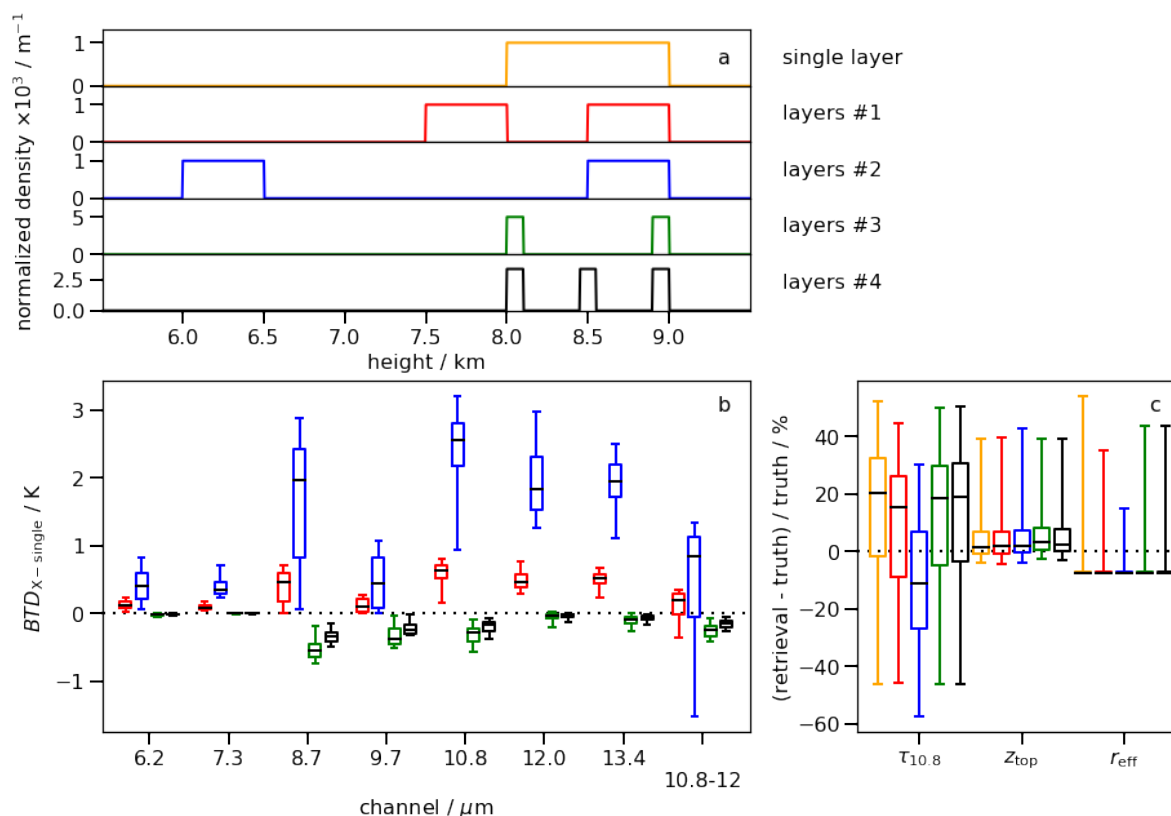


Figure 6. Analysis of the impact of different ash layerings with (a) the normalized vertical cloud profiles for multiple layers, (b) the $BTDX - \text{single}$ of different channels between a single mass layer and the layerings in (a), (c) the relative differences in the retrievals of $\tau_{10.8}$, z_{top} and r_{eff} compared to the true values for the different layerings in (a); for each setup 500 simulations are averaged (see text); the boxplot shows the median, first and third quartile (box) and the 5th and 95th percentile (whiskers).

3.3. Geometrical Ash Cloud Thickness

Although the geometrical cloud thickness is varied within the training data set, it is not retrieved explicitly by the ANNs, and it is not clear whether they are able to derive this information internally as a side product. Here we quantify the impact of this property with respect to the brightness temperature and the retrieval results. Therefore, layers with thicknesses of 0.5 km, 1 km, 2 km and 3 km are considered, with $z_{\text{top}} = 9$ km and $m_{\text{col}} = 10 \text{ g m}^{-2}$ kept constant. The difference compared to the 1 km thick case are considered, Figure 8a. As expected, thicker clouds have higher brightness temperatures as more mass is in lower and warmer parts of the troposphere; the thinner layer leads to lower brightness temperatures. The geometrical cloud thickness introduces a variability of the brightness temperatures of -1 to 4 K. Figure 8b shows the results of the retrievals. Whereas the impact of the cloud thickness is generally negligible for r_{eff} , an increased thickness (and, therefore, decreased cloud base height) leads to smaller retrievals for $\tau_{10.8}$ and z_{top} . The influence on the latter is only small (between -4% and 8%), but the former can exhibit differences of more than 20% comparing cloud geometrical thicknesses of 0.5 km and 3 km.

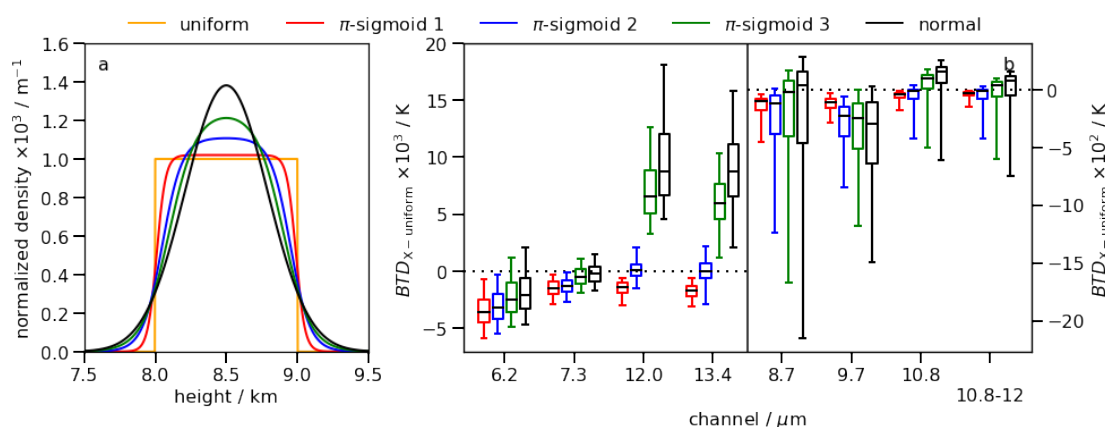


Figure 7. Analysis of the impact of different ash profiles with the same mean and standard deviation with respect to the height of the center of the cloud, with (a) the different profiles and (b) the BTD of different channels between a uniform vertical mass distribution and the other distributions in (a); for each setup 500 simulations are averaged (see text); the boxplot shows the median, first and third quartile (box) and the 5th and 95th percentile (whiskers).

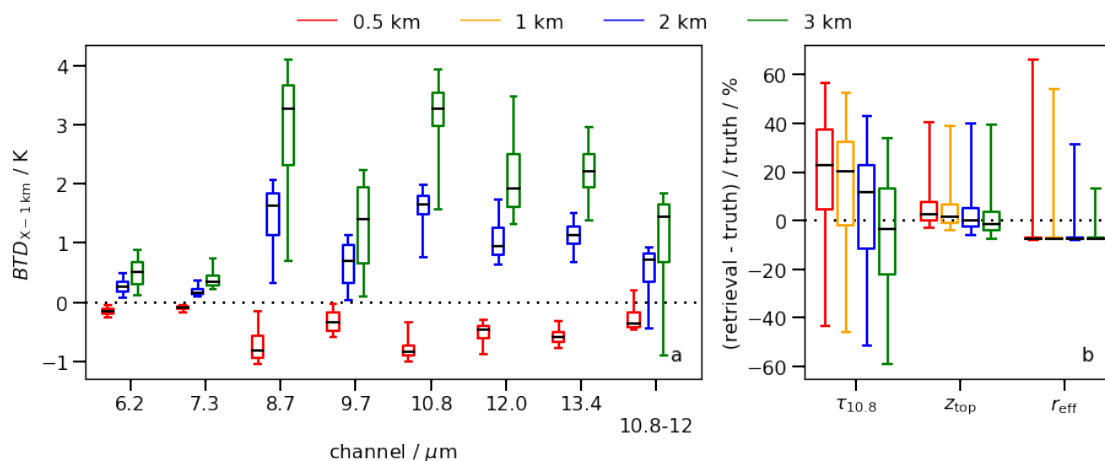


Figure 8. Analysis of the impact of different geometrical ash layer thicknesses, with (a) the BTD of different channels between a layer with a thickness of 1 km and the other thicknesses, (b) the relative differences in the retrievals of $\tau_{10.8}$, z_{top} and r_{eff} compared to the true values for the different layer thicknesses; for each setup 500 simulations are averaged (see text); the boxplot shows the median, first and third quartile (box) and the 5th and 95th percentile (whiskers).

Comparing different volcanic ash clouds, geometrical layer thickness and multi-layering represent the largest sources of error. The uncertainties introduced by those properties with respect to the brightness temperatures are larger than the instrumental noise [19] and are of the order of 30% for $\tau_{10.8}$, 5% for z_{top} and negligible for r_{eff} . The shape of a single layer profile is negligible.

4. Comparisons with Independent Measurements

To prove the applicability of the retrievals to real data, we compare our results with other in situ and remote sensing measurements as well as the outcome of the predecessor VADUGS for selected scenes.

4.1. Puyehue-Cordón Caulle Eruption (2011)

Lidar measurements represent an excellent source for comparison since they provide accurate estimates for z_{top} and the vertical profile. Here we use data from the CALIPSO/CALIOP [20] version 4.10 level 2 aerosol products, which include information about volcanic ash layers in the stratosphere [21]. The extinction profiles of those

layers were used to calculate their optical depth at 532 nm, which was converted to m_{col} using a mass extinction coefficient of $690 \text{ m}^2 \text{ kg}^{-1}$ [22,23]. z_{top} corresponds to the top of the uppermost ash layer. Only ash samples with extinction quality control flag zero (initial lidar ratio resulted in stable extinction retrievals, i.e., “unconstrained” retrievals) or one (when the lidar ratio could be inferred directly from the data, i.e., “constrained” retrievals) were used. For the unconstrained retrievals, a lidar ratio of 58 sr (median of the directly retrieved lidar ratios) was used to update the ash optical depth to correct for the low bias resulting from the use of the default lidar ratio of 44 sr for ash in CALIPSO version 4.10. The final data set has a horizontal resolution of 5 km and a vertical resolution of 60 m.

The six scenes in consideration are listed in Table 3 and sketched in Figure 9; data are plotted in Figure 10 (blue line). They show volcanic ash clouds from the Puyehue-Cordón-Caulle eruption, starting at 4 June 2011. The flyovers took place above the southern Atlantic ocean between 15 June 2011 and 18 June 2011 during day and night. z_{top} was 10–15 km and m_{col} reached up to 1.5 g m^{-2} ; note that the uncertainty of the latter is roughly a factor of 2, see Figure 10. r_{eff} is not derived from CALIOP data, but Bignami et al. [24] used MODIS data to retrieve mean values of 4–6 μm within 300 km from the volcano. Due to sedimentation processes r_{eff} should be smaller in the scenes considered here. Ishimoto et al. [25] retrieved from IASI spectra at distances of $\sim 1000 \text{ km}$ for most samples $r_{\text{eff}} < 0.5 \mu\text{m}$. The silica content was around 70 wt.% or slightly below [14,26,27].

Table 3. Investigated CALIPSO flyovers in June 2011 with timespan, coordinates of start and endpoint, number of samples and track number according to Figure 9.

Day	Start Time/UTC	End Time/UTC	Start Coordinates	End Coordinates	Samples	Track Number
15 June 2011	18:30	18:40	−55.7°N, −59.8°E	−40.4°N, −66.0°E	300	1
16 June 2011	15:51	16:05	−48.8°N, −24.4°E	−39.3°N, −27.7°E	162	2
16 June 2011	17:29	17:43	−60.9°N, −42.7°E	−44.6°N, −50.7°E	187	4
17 June 2011	03:00	03:13	−40.6°N, −27.7°E	−62.0°N, −37.8°E	251	5
17 June 2011	14:55	15:10	−44.4°N, −12.1°E	−37.3°N, −14.4°E	82	3
18 June 2011	02:04	02:18	−35.4°N, −12.2°E	−64.4°N, −25.9°E	199	6

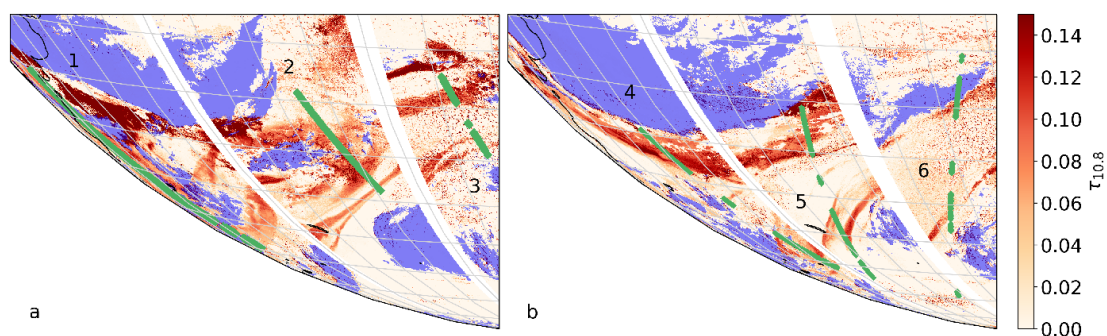


Figure 9. Overview of the CALIPSO transits (green) as described and numbered in Table 3 with the corresponding VACOS retrievals of $\tau_{10.8}$ (red colors) and the cirrus flag of CiPS (blue); Figure (a,b) are made up of three stripes each, showing the VACOS and CiPS retrievals at the times of the corresponding CALIPSO transits; thus, the retrievals of the three stripes in each plot correspond to three different times.

We compare the CALIOP measurements with the results of the ANNs applied to SEVIRI images at close times. The coordinates given by CALIPSO are corrected such that the light path of the SEVIRI measurements penetrates the top of the ash cloud (parallax correction). The regression retrievals are shown on a single pixel basis (faint red), and after a mean filter of 5×5 pixels was applied (red); the classification result is shown only on a single pixel basis. $\tau_{10.8}$ is converted to m_{col} using a mean mass extinction coefficient at $10.8 \mu\text{m}$ of $152 \text{ m}^2 \text{ kg}^{-1}$, which holds for silica contents of 70 wt.% and $r_{\text{eff}} = 0.6 \mu\text{m}$ [2].

Additionally, the retrievals of volcanic ash by VADUGS and of cirrus clouds by CiPS [9,10] are shown in Figure 10. The MAPE and MPE for the retrievals of m_{col} and z_{top} are given in Table 4, the PODs using m_{col} and P_{ash} and different thresholds are given in Figure 11.

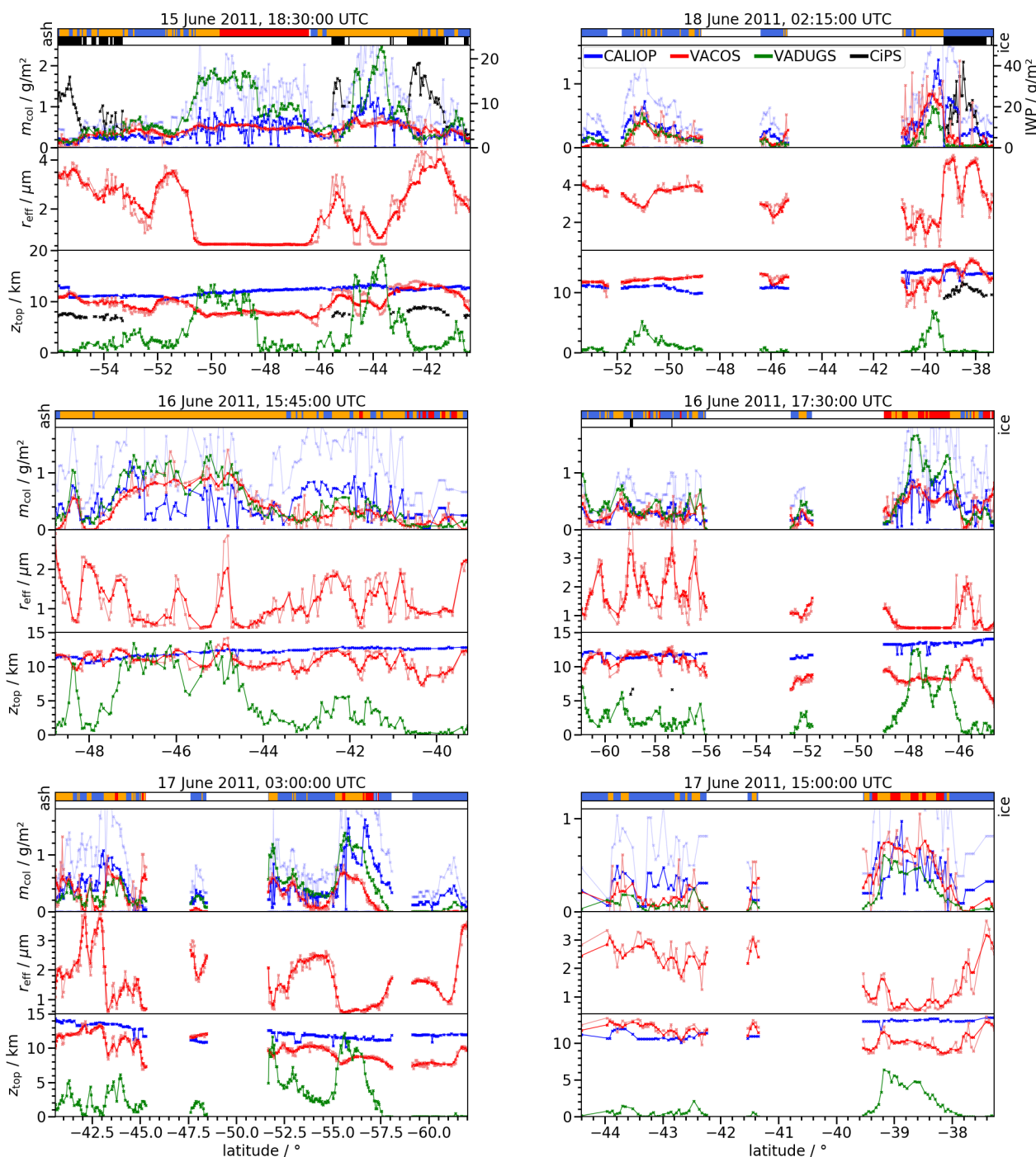


Figure 10. Pixel classification (*ash*) and retrievals of m_{col} (derived from $\tau_{10.8}$), r_{eff} and z_{top} as determined using CALIOP (blue), VADUGS (green) and VACOS (red) for volcanic ash clouds of the Puyehue-Cordón Caulle eruption 2011 above the Atlantic on 15–18 June 2011, see Table 3; the times above the plots indicate the start of measurement of the SEVIRI image used; the VACOS regression results are averaged on 5×5 pixels with the 1×1 pixel result shown in faint red, whereas the for the classification the single pixel result is shown; the upper uncertainty of m_{col} by CALIOP is indicated in faint blue; the ash classification shows: clear skies (green), meteorological clouds (blue), volcanic ash (red), ash and meteorological clouds (orange); a cirrus cloud flag (*ice*) and, if applicable, the ice water path (IWP) are derived using CiPS (black); the cirrus flag shows: cirrus (black), no cirrus (white); the latitude refers to the position of CALIPSO; note that vertical axes are scaled differently in the plots.

Table 4. Comparison of the retrievals VADUGS and VACOS against CALIOP data; different subsets depending on m'_{col} from CALIOP are considered; the MAPE and MPE for the retrieval of m_{col} and z_{top} are calculated; the retrieval is analyzed after application of a 3×3 pixels (px) and 5×5 px uniform filter.

Algorithm	m_{col}		z_{top}	
	MAPE	MPE	MAPE	MPE
full data set (1181 samples)				
VADUGS	123%	70%	75%	−74%
VACOS (1 px)	190%	93%	19%	−14%
VACOS (9 px)	111%	43%	18%	−14%
VACOS (25 px)	112%	49%	18%	−14%
only $m'_{\text{col}} \geq 0.2 \text{ g m}^{-2}$ (875 samples)				
VADUGS	62%	12%	71%	−70%
VACOS (1 px)	56%	−24%	18%	−14%
VACOS (9 px)	47%	−20%	18%	−14%
VACOS (25 px)	45%	−20%	18%	−14%

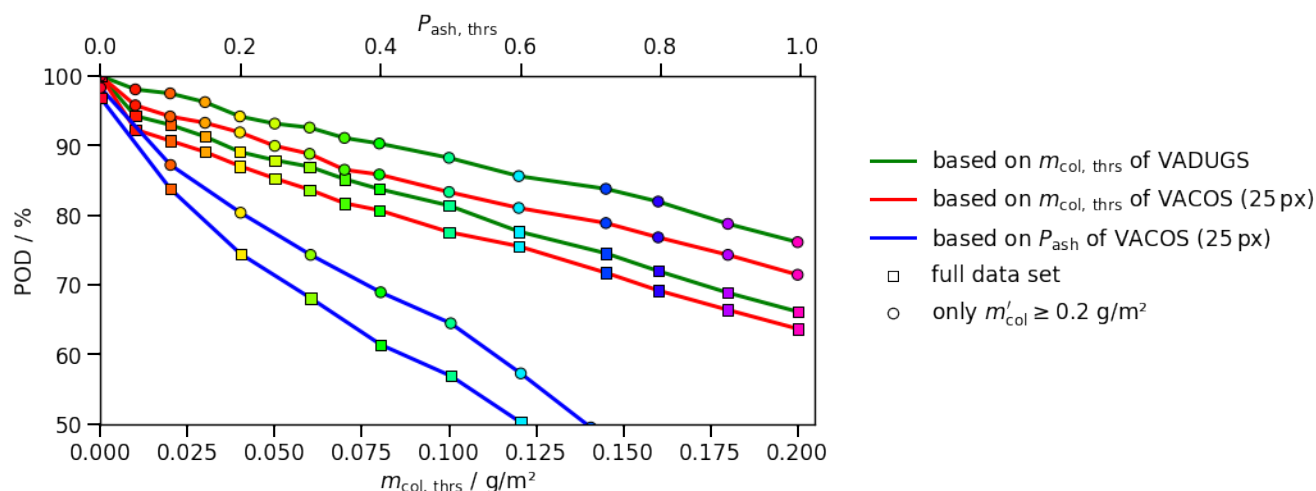


Figure 11. Estimations of the POD of volcanic ash using m_{col} and P_{ash} of VACOS (connected by red and blue lines, respectively) and m_{col} of VADUGS (green line), and a corresponding threshold; different subsets of the test data (Table 3), are used and encoded in the marker type, i.e., the full data set (square), only samples with target $m'_{\text{col}} \geq 0.2 \text{ g m}^{-2}$ (circle).

Again the mass load due to CALIOP (i.e., the “true” value) is denoted m'_{col} . For $m'_{\text{col}} \geq 0.2 \text{ g m}^{-2}$, the retrieval of m_{col} using VADUGS has a MAPE of 62% and a MPE of 12%, indicating a slight overestimation; this is visible in Figure 10 for 15 June 2011 at 18:30 UTC or 16 June 2011 at 17:30 UTC, although the VADUGS results are mostly within the uncertainty interval of the CALIOP data. VACOS has a smaller MAPE of 56%, but underestimates m_{col} due to a MPE of −24%. Remember that VACOS is more complex than VADUGS, consisting of four instead of only one ANN, using more input features, with each ANN having three compared to just a single hidden layer (although with 100 instead of 600 neurons per hidden layer) [1,2]. As a consequence, VACOS has significantly more trainable parameters and has the potential to learn more complex functions, such that their pixelwise application can lead to rather abrupt jumps in the retrievals; this can be seen, for instance, at 18 June 2011, 02:15 UTC in Figure 10 and the corresponding track 6 in Figure 9b. Thus, the VACOS retrievals are also calculated after the application of a 3×3 and a 5×5 pixels uniform averaging, which leads to a further reduction of MAPE (47% and 45%, respectively) and MPE (−20% each). Note that this MAPE is similar to the values found for the simulated test data. Whereas the reduction in the errors is of the order of 10% when comparing the unfiltered results with the results after an averaging over 3×3 pixels, the further decrease is only on the order of a few percent when considering

5×5 pixels. Note that averaging over even larger areas might worsen the retrieval of fine structured ash clouds, e.g., thin plumes close to the volcanic vent; therefore, this is not done here. Considering the full CALIOP data set leads to significantly higher errors than for the subset with $m'_{\text{col}} \geq 0.2 \text{ g m}^{-2}$, showing that the retrieval of thin ash layers is much harder. VADUGS has a MAPE of 123% and a MPE of 70%. VACOS has higher errors before averaging the results, but performs better after averaging over 3×3 pixels, with a MAPE of 111% and a MPE of 43%. Averaging over 5×5 pixels slightly increases the errors. The reason might be that small m_{col} are often located at the edges of ash clouds, where the averaging over larger areas includes ash-free pixels. The retrieval of z_{top} is mostly independent of the subset and the averaging for both VADUGS and VACOS. MAPE and absolute MPE of VADUGS are 70–75%. Although VADUGS is able to retrieve the correct heights in some cases (especially when $m_{\text{col}} \geq 1 \text{ g m}^{-2}$), it often strongly underestimates z_{top} ; compare, for instance, 16 June 2011 at 15:45 UTC in Figure 10. VACOS retrieves z_{top} with a MAPE $< 20\%$, but also slightly underestimates the true height as indicated by a MPE of -14% . The retrieved r_{eff} exhibits large variations with values between $0.6 \mu\text{m}$ (the lower end of the training data regime [2]) and $4 \mu\text{m}$; thus, they are smaller than the values retrieved by Bignami et al. [24] close to the volcano, but significantly larger than the results by Ishimoto et al. [25]. Note that the lowest r_{eff} are retrieved for the highest m_{col} (e.g., at 16 June 2011, 17:30 UTC). As the errors of r_{eff} decrease with increasing $\tau_{10.8}$ (as shown in Section 2), the retrieved r_{eff} is most reliable for thick clouds. The small r_{eff} also supports the choice of the mass extinction coefficient at $10.8 \mu\text{m}$ above.

The classification ANN indicates the presence of volcanic ash especially for thick ash layers, whereas thinner layers are often misclassified as meteorological clouds. The POD using P_{ash} of the resulting binary ash flag and using m_{col} of VADUGS and VACOS (compare Section 2.3) are given for different thresholds in Figure 11. For VACOS, the averages over 5×5 pixels are used for m_{col} and the single pixel result for P_{ash} . VACOS performs slightly worse than VADUGS with respect to volcanic ash detection using m_{col} . Setting $m_{\text{col,thr}} = 0.2 \text{ g m}^{-2}$ leads to a POD of 65–75% for VACOS; for $m_{\text{col,thr}} = 0.1 \text{ g m}^{-2}$, the POD is around 80%. Both values are significantly smaller than for the simulated test data set, which indicates that there are differences between the simulated data set and data collected in reality. However, note that the different performances of VADUGS and VACOS can be explained in part by the retrievals MPEs: VADUGS overestimates m_{col} whereas VACOS underestimates it. Correcting this would lead to a decrease of the POD of VADUGS and to an increase for VACOS. Furthermore, the assumed mass extinction coefficient at $10.8 \mu\text{m}$ for the transformation of $\tau_{10.8}$ to m_{col} has an impact for VACOS: using a smaller value increases m_{col} due to VACOS and consequently also the POD. The performance of the binary ash flag is significantly lower here as compared to the simulated test data and the results using the retrieved m_{col} , independently of the threshold; $P_{\text{ash,thr}} = 0.5$ leads to a POD around 60%. This indicates that the use of m_{col} for detection is more reliable than the classification ANN in this situation. The FAR has not been quantified here, but Figure 9 shows mostly large-scale structures representing the ash clouds, and only in some scenes are tiny patches with significant $\tau_{10.8}$ which are not connected to the ash clouds and might be false detections. Thus, we assume the FAR to be reasonably low.

Two scenes, 15 June 2011 at 18:30 UTC and 18 June 2011 at 02:15 UTC, contained cirrus clouds underneath the ash layer, with ice water paths up to 20 g m^{-2} and 40 g m^{-2} , respectively. In both cases the cirrus has a moderate impact on the retrieved z_{top} , but the retrieved r_{eff} is significantly increased in the presence of the cirrus clouds; meanwhile, the retrieved $\tau_{10.8}$ drops to zero, especially at 18 June 2011, 02:15 UTC. This shows that it seems recommendable to always evaluate VACOS results alongside cirrus retrievals.

4.2. Eyjafjallajökull Ash Cloud (17 May 2010)

Following the 2010 eruption of the Eyjafjallajökull, the Deutsches Zentrum für Luft- und Raumfahrt (DLR, the German aerospace center) and the Facility for Airborne Atmospheric Measurements, United Kingdom (FAAM) independently performed airborne

in situ and lidar measurements of the volcanic ash clouds [15,16,28]. Here we consider measurements from 17 May 2010, when both aircraft simultaneously investigated the same ash cloud above the North Sea [17]. The VACOS retrieval of $\tau_{10.8}$ and the cirrus mask of CiPS are given for an example scene, Figure 12b, showing that the ash and ice clouds are well separated. VACOS can be compared to volcanic ash detections by Schumann et al. [15] for a scene two hours later, using $BT D_{10.8-12}$ with a threshold of -1 K for detection and applying a low-pass filter (Figure 15 in [15]). Both methods exhibit similar distributions of the optically thickest ash clouds, e.g., patches above the North Sea, west of Norway and above Iceland. However, VACOS finds additional, extended ash clouds with $\tau_{10.8} \leq \sim 0.05$ above Germany and Norway. The ash cloud investigated by DLR and FAAM as well as the flight trajectories of the two aircraft are sketched in Figure 12a.

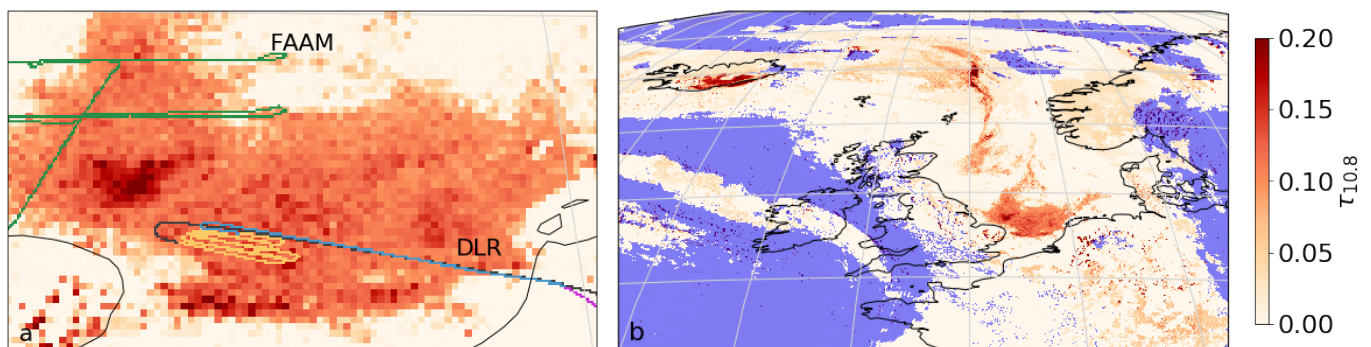


Figure 12. VACOS retrieval of $\tau_{10.8}$ at 17 May 2010, 16:00 UTC; shown is (a) the North Sea between the Netherlands and England, and (b) north-western Europe up to Iceland; blue areas are covered by cirrus clouds according to CiPS [9]; panel (a) also shows the flight tracks of the DLR Falcon aircraft (with different parts of the track in black, yellow, blue and violet) and the FAAM aircraft (green) [15,16].

Schumann et al. [15] provided in situ measurements with a time resolution of 1 s for the altitude and 10 s for the mass volume concentration. To derive the latter they made two different assumptions on the refractive index of the volcanic ash (case L: $1.59 + 0.0i$, case M: $1.59 + 0.004i$; at 630 nm), leading to an upper and a lower estimate of the mass concentration. A third case (H: $1.59 + 0.008i$) was evaluated by Schumann et al. [15], but based on their full analysis they expected the true value to be between case L and M. Ball et al. [29] measured for the refractive index of Eyjafjallajökull ash at a wavelength of 650 nm a real part of 1.554 ± 0.01 and an imaginary part of 0.00085 ± 0.00069 , which would support case L. Schumann et al. [15] pointed out that the imaginary part of the refractive index was the major source of uncertainty. The effective radius was estimated to be $0.65\text{--}1.05\text{ }\mu\text{m}$ [15]. Considering altitude and mass concentration shows that the aircraft enters the ash cloud from above at 15:50 UTC, reaches a minimum altitude around 16:35 UTC (yellow in Figure 12), and rises afterwards until it leaves the ash cloud again at ca. 17:00 UTC (blue in Figure 12), see Figure 13. Data of the two transits (from above and from below) are treated separately; the corresponding measured vertical mass profiles as well as their linearly interpolated and over 200 m averaged profiles are shown in Figure 14. Although different parts of the ash cloud are probed during the two flight legs, the vertical mass profiles are relatively similar, indicating a horizontally homogeneous ash distribution (over distances in the order of ~ 50 km). However, vertically the ash cloud has a highly variable mass profile. Ash cloud top and bottom height are roughly at 6.3 km and 3.3 km, respectively. Integrating the vertical mass profiles of Figure 14 and averaging m_{col} of the two flight legs gives 0.52 g m^{-2} and 2.13 g m^{-2} for the lower (L) and upper (M) estimate, respectively. The DLR data are shown in the left panel of Figure 13.

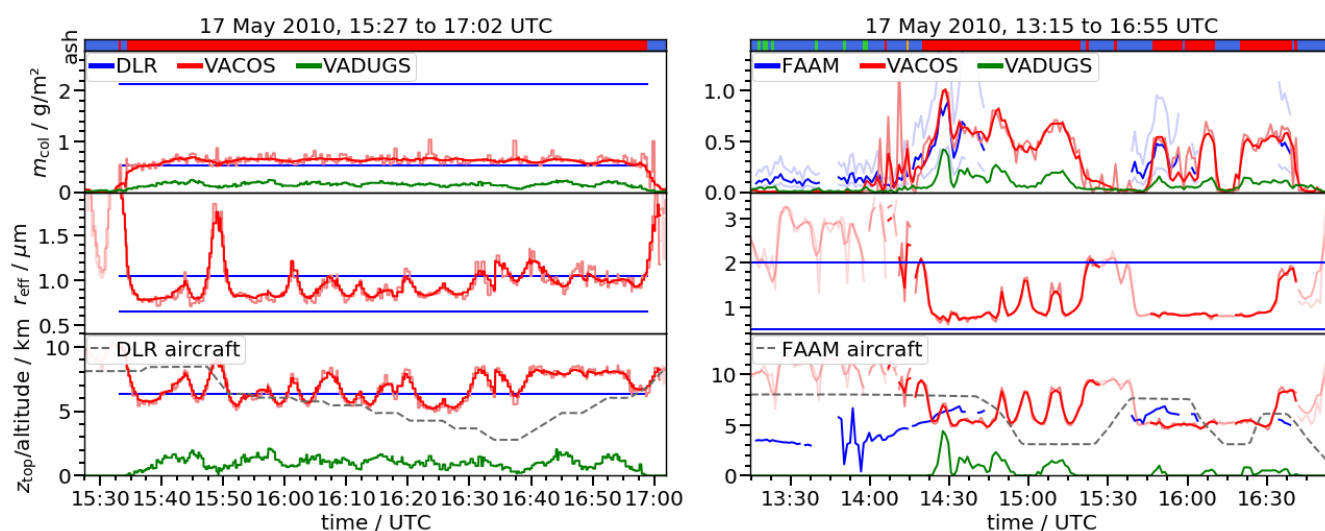


Figure 13. Pixel classification (*ash*) and retrievals of m_{col} (derived from $\tau_{10.8}$), r_{eff} and z_{top} as measured in situ by DLR (blue, left panel [15]), derived using an airborne lidar by FAAM (blue, right panel [16]), VADUGS (green) and VACOS (red) for an Eyjafjallajökull ash cloud above the North Sea at 17 May 2010; the VACOS regression results are averaged on 5×5 pixels with the 1×1 pixel result plotted in faint red, whereas only the single pixel result is shown for the classification; also VACOS retrievals of r_{eff} and z_{top} for $m_{col} < 0.1 \text{ g m}^{-2}$ are plotted in faint red; for the in situ data, only mean values or upper and lower estimates are given; for the lidar-derived data, the upper and lower uncertainty of m_{col} is indicated in faint blue; the ash classification shows: clear skies (green), meteorological clouds (blue), volcanic ash (red), ash and meteorological clouds (orange); the altitudes of the aircrafts are given (grey dashed).

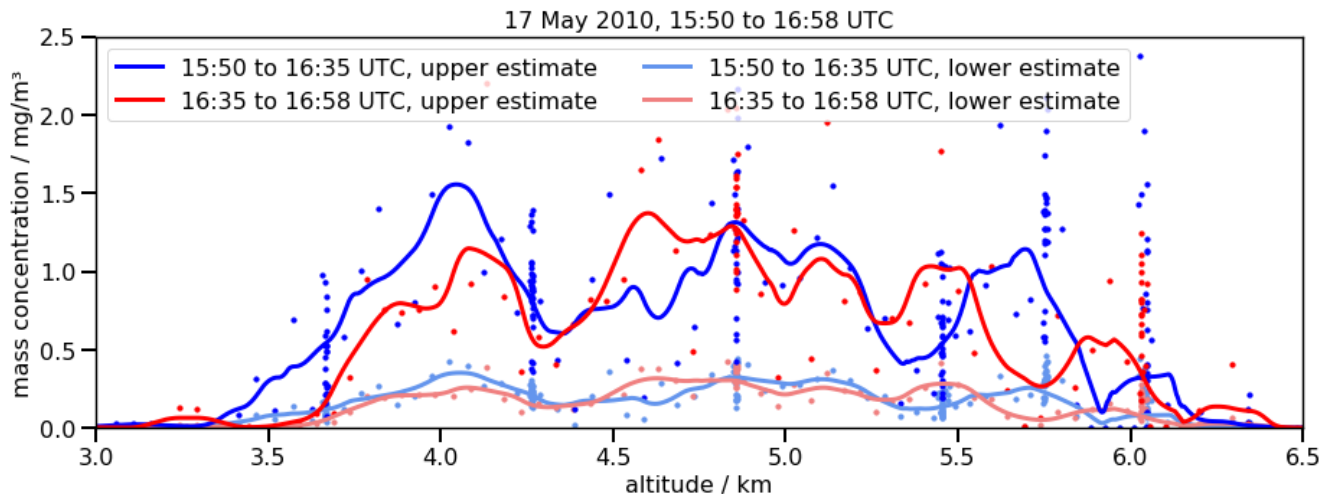


Figure 14. Vertical mass profiles of the volcanic ash cloud as measured by Schumann et al. [15] during two flight legs (see text) indicated by the color (blue, red); based on different assumptions, there is an upper estimate (M, bold colors) and a lower estimate (L, faint colors) of the in situ measurements; 10 s averages are given as dots, derived 200 m means are shown as lines.

From Marenco et al. [16] the lidar-derived mass profiles are used, which have a vertical resolution of 45 m and a temporal resolution of 60 s, corresponding to a horizontal resolution of $9 \pm 2 \text{ km}$ at typical aircraft speeds. The uncertainty of the masses is given by a factor of two [16]. We calculate m_{col} by integrating the vertical mass profile, whereas we define z_{top} as the height where the mass volume concentration (median over 315 m, i.e., 7 height levels) exceeds $50 \mu\text{g m}^{-3}$; averaging kernel and mass concentration threshold are based on visual inspection of the mass profiles. In situ measurements performed in April and May 2010 by the FAAM led to r_{eff} of ca. $0.5\text{--}2 \mu\text{m}$ [28]. The FAAM data for

the flight at 17 May 2010 are shown in the right panel in Figure 13. Just as the DLR, the FAAM aircraft dipped into the ash cloud several times; to avoid incomplete mass profiles, measurements of m_{col} and z_{top} are discarded if the FAAM aircraft is <500 m above z_{top} or if the aircraft is below 5 km.

Apart from different instrumentation and processing for the in situ measurements [17], the flights by DLR and FAAM followed fundamentally different strategies: the DLR aircraft remained in an self-contained area well within the ash cloud, hence the relatively constant m_{col} , whereas the FAAM entered and left the ash-containing area multiple times, see Figure 12a. Still, the data from both measurements shows some similarities, Figure 13. The lower estimate of m_{col} from DLR is generally in good agreement with the maximum best estimate m_{col} due to the FAAM, whereas the upper estimate from DLR and the maximum of the uncertainty of m_{col} by the FAAM are both around 2 g m^{-2} . The lower estimate of r_{eff} is similar for DLR and FAAM, but the upper estimate differs by roughly $1 \mu\text{m}$; differences in the ash particle size distributions were also reported by Turnbull et al. [17]. z_{top} from DLR agrees with the largest estimate derived from FAAM data, but the latter varies between roughly 4 km and 7 km, again indicating that a more diverse area of the ash cloud was sampled.

For the VACOS and VADUGS retrievals, a parallax correction is implemented, i.e., coordinates are considered such that the light path crosses the coordinates of the DLR and FAAM aircraft at a height of 6 km and 5 km, respectively, and penetrate the observed ash clouds. The heights correspond roughly to the measured z_{top} of DLR/FAAM. The SEVIRI images with a time resolution of 15 min are processed and the temporally closest image is chosen for each measurement; the SEVIRI line acquisition time is considered. The FAAM's mean aircraft velocity of $\sim 146 \text{ m s}^{-1}$ (on ground) and a SEVIRI pixel size of roughly 6.8 km times 3.2 km (at 54°N , 1.5°E) indicate that it takes the aircraft generally about 47 s to cross a pixel in north–south direction and 22 s in east–west direction. As the FAAM data has a time resolution of 60 s, a minimum averaging of 1×3 SEVIRI pixels is necessary to compare the satellite retrievals to the FAAM data, assuming the aircraft moves only in east–west direction. Based on the results of Section 4.1, an averaging over 5×5 pixels is considered here. As r_{eff} due to DLR and FAAM data is around $1 \mu\text{m}$, a mass extinction coefficient at $10.8 \mu\text{m}$ of $200 \text{ m}^2 \text{ kg}^{-1}$ is assumed [2]; note that this conversion factor is applied in the rest of this work when dealing with Eyjafjallajökull ash clouds. VACOS retrievals of r_{eff} and z_{top} are shown in faint red if $m_{\text{col}} < 0.1 \text{ g m}^{-2}$, as the former are ill-defined in the absence of ash clouds. The resulting retrieval data is given in Figure 13.

Considering the DLR case (left panel in Figure 13) we find that the VACOS m_{col} lies at the lower end of the uncertainty interval of the DLR measurement with relatively low variability. z_{top} varies between 5.5 km and 8.5 km, which includes the value found from the DLR data. For the most parts, r_{eff} lies well between the estimates by Schumann et al. [15]. The classification ANN flags the whole ash cloud correctly.

The FAAM measurements (right panel in Figure 13) and the VACOS retrievals for m_{col} , z_{top} and r_{eff} are generally in agreement. VACOS slightly underestimates the mass load with $m_{\text{col}} \approx 0 \text{ g m}^{-2}$ compared to the FAAM value of $\sim 0.1 \text{ g m}^{-2}$ before 14:00 UTC (and consequently derives too high estimates for r_{eff} and z_{top}), but retrieves similar m_{col} around 0.6 g m^{-2} as the FAAM later on. VACOS retrieved r_{eff} and z_{top} show plateaus around $0.9 \mu\text{m}$ and 5 km, respectively, at 14:30 to 14:45 UTC as well as 15:45 to 16:30 UTC. At other times, e.g., around 14:20 or 15:20 UTC, there are coincident increases of r_{eff} and z_{top} , which can be attributed at least partially to thin ash clouds with $m_{\text{col}} < 0.3 \text{ g m}^{-2}$ at the edges of the ash clouds; thus, these retrievals can be assumed to be unphysical. Other peaks in r_{eff} and z_{top} , e.g., around 15:00 UTC, appear although there is a significant amount of volcanic ash (i.e., VACOS retrieves $m_{\text{col}} \geq 0.4 \text{ g m}^{-2}$). The classification correctly classifies the major parts of the ash encounters.

Although the comparisons of VADUGS retrievals with the CALIOP results in the case of the Puyehue-Cordón Caulle ash cloud showed a good agreement, VADUGS seems

hardly applicable in the present case: m_{col} and z_{top} are strongly underestimated in both comparisons and at all times.

4.3. Eyjafjallajökull Ash Plume at Vent (11 May 2010)

The concentrations that have been compared up to now are rather low. Here the retrieval ability closer to the vent is investigated. Researchers from the University of Iceland used a piston-driven aircraft to probe the outer parts of the ash plume of Eyjafjallajökull in 2010 at distances of 45–60 km from the vent [30]. Plume heights of 3–4.3 km have been reported [30,31], and at 11 May 2010, mass volume concentrations of $0.5\text{--}2\text{ mg m}^{-3}$ [30]. Assuming a geometrical thickness of 1–2 km, this corresponds to m_{col} being $0.5\text{--}4\text{ g m}^{-2}$ for the fringes of the ash plume; in the center of the plume the values should obviously be higher. Note that Weber et al. [30] reported the boundary of the ash plume to be inhomogeneous, with ash cloud puffs of diameter 0.5–2 km. These small scale structures are not resolvable by SEVIRI [2].

As an example we consider a scene at 11 May 2010, 14:00 UTC, Figure 15. The Eyjafjallajökull ash plume moves southwards from the vent and is surrounded by different cloud fields (a). The cirrus cloud retrieval CiPS [9,10] shows ice water clouds collocated with the ash plume up to some hundred kilometers from the vent. With roughly $1\text{--}6\text{ g m}^{-2}$, the ice water path (b) is rather low (e.g., compared to the retrievals in Figure 10), and the cirrus clouds are located at 9–11 km height (c), i.e., above the heights given by Weber et al. [30]. Weber et al. [30] also report that the top of the ash plume above the sea appeared white, whereas it was darker below, and significant ice contents have been found in volcanic ash plumes before [32–34]. The additional ice content might spoil the volcanic ash retrieval in this scene (compare Section 4.1). Note that although it was shown that aerosol layers below cirrus clouds have only a small impact on CiPS [10], the volcanic ash plume in this scene is by no means negligible (f). Therefore, the CiPS retrievals have to be treated with caution as well. The ash plume is detected by VADUGS (d), but with $m_{\text{col}} \approx 0.03\text{ g m}^{-2}$ it underestimates the mass load by one to two orders of magnitude compared to Weber et al. [30]. The classification due to VACOS (e) shows the (mostly pure) ash plume moves southwards and then bends eastwards. A visual comparison of the classification results with (a) indicates that too few pixels might be classified as clear sky (green). The VACOS retrievals of m_{col} (f), z_{top} (g) and r_{eff} (h) are shown for pixels with $\tau_{10.8} > 0.02$. m_{col} (f) shows the plume with decreasing values downwind. The upper end and the center of the ash plume show values around 3 g m^{-2} and larger, the edges values of the order 1 g m^{-2} ; this is in agreement with the in situ measurements by Weber et al. [30]. Downwind the plume bends eastwards. The ash has dispersed and m_{col} is ca. $0.5\text{--}1\text{ g m}^{-2}$. The retrieved z_{top} (g) has values of 10–12 km close to the vent, which is much larger than the literature values. When the plume bends eastward, z_{top} retrievals drop to 5–6 km. Similarly, we find large r_{eff} (h) around $3\text{ }\mu\text{m}$ close to the volcano and then a sudden drop to values around $1\text{ }\mu\text{m}$. In all three cases the sudden change in the retrievals happens at the edge of the retrieved cirrus cloud, compare (b) and (f), indicating that the presence of the cirrus cloud might lead to the overestimation of m_{col} , z_{top} and r_{eff} . Apart from the impact by the cirrus cloud, the retrievals also have some inherent limitations that might lead to inaccuracies close to the vent: the ash properties used in the training data correspond rather to those of aged ash clouds than to fresh ash (e.g., only small particles without porosity are considered) and typical gas emissions (e.g., water vapor, sulfur dioxide) have not been included [2].

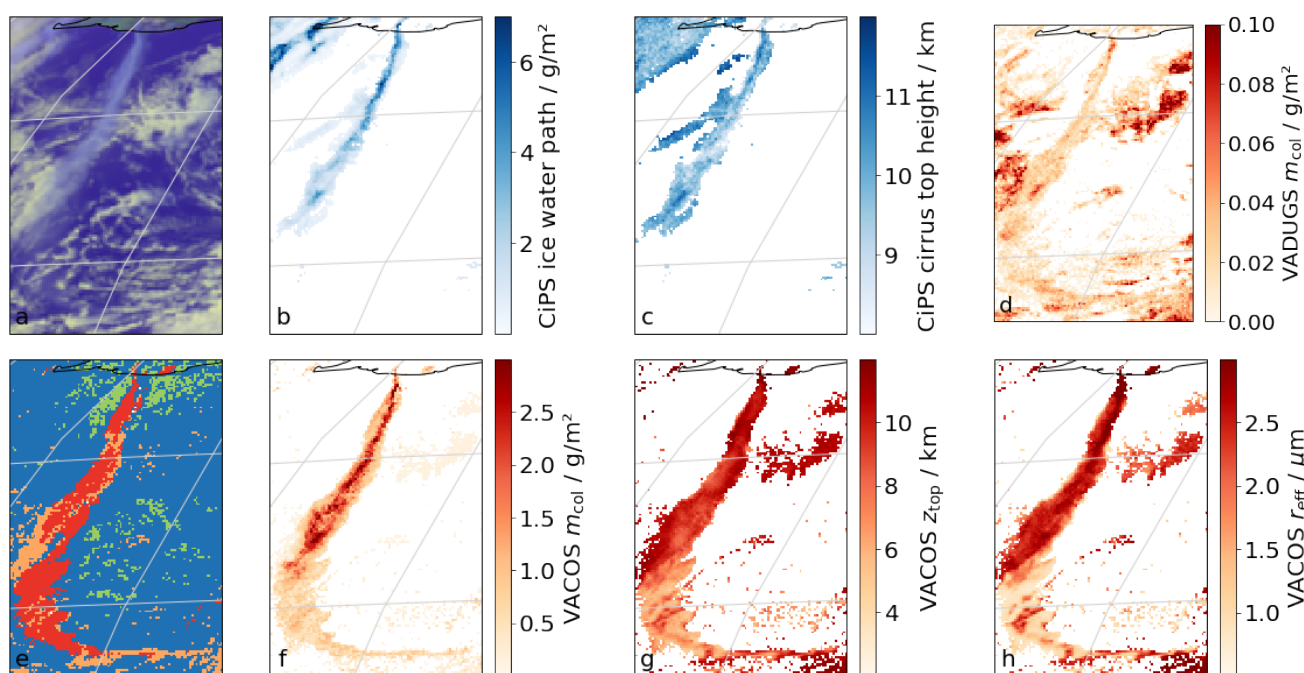


Figure 15. Retrieval results for the Eyjafjallajökull eruption at 11 May 2010, 14:00 UTC with Iceland outlined in the top of the plots; (a) false color overview composite from SEVIRI data, from CiPS (b) cirrus ice water path and (c) cloud top height, (d) m_{col} from VADUGS, and from VACOS (e) the classification (color code as in Figure 10), (f) m_{col} , (g) z_{top} and (h) r_{eff} .

5. Comparison with a Model Ensemble

To check the general performance of VACOS on large scales, the retrievals are compared with the results of a volcanic ash transport and dispersion model. Note that the model is an approximation of the reality as, for instance, inaccuracies in the volcanic ash source term or the meteorological conditions are transferred to the ash distribution. Thus, the scope of this section is to quantify the agreement of the spatial distribution of volcanic ash clouds and to compare the order of magnitude of m_{col} .

Here the multi model multi source term ensemble by Plu et al. [35] is used. They simulated the Eyjafjallajökull eruption at 13–19 May 2010 and the area from Iceland in the north-west to Italy in the south-east. We consider scenes every six hours. Four models were used: the atmospheric Lagrangian transport model FLEXPART, the Eulerian chemical transport model MATCH and the chemical transport models MOCAGE and WRF-Chem, considering in varying ways phenomena such as atmospheric transport and mixing, gravitational settling, wet and dry deposition, and chemical reactions. An a posteriori source term is used, determined using a FLEXPART-based optimal estimation model and estimates of m_{col} from satellite data, as well as upper and lower bounds for the source term based on the uncertainties of the optimal estimation result. All models spin-up for at least 3 days, their results are vertically integrated and averaged on a $0.2^\circ \times 0.2^\circ$ grid, as only the large-scale distribution of ash is of interest here. The ensemble result m'_{col} is the median of all simulations.

VACOS and VADUGS retrievals of m_{col} are averaged on the same grid. Instead of the simple uniform averaging of the retrieval results performed in Section 4, the following accumulation rule is applied to alleviate the impact of (a) the different spatial resolutions, (b) possible temporal and spatial shifts between model and satellite retrieval and (c) false satellite detections: a grid cell at time t is assumed to be ash-contaminated if at least a fraction f_{thrs} of all pixels within the cell and within the time $[t - 15 \text{ min}, t + 15 \text{ min}]$ exceeds a threshold of $m_{col,thrs}$; then the estimate of m_{col} is the mean of this fraction of the pixels. For VADUGS, $f_{thrs} = 0.5$ and $m_{col,thrs} = 0.1 \text{ g m}^{-2}$ is used. For VACOS, three different settings are applied: the same as for VADUGS; $f_{thrs} = 0.5$ and $m_{col,thrs} = 0.2 \text{ g m}^{-2}$;

$f_{\text{thrs}} = 0.9$ and $m_{\text{col, thrs}} = 0.2 \text{ g m}^{-2}$. Grid cells covered by cirrus clouds according to the retrieval COCS [36] are discarded.

Three example scenes are shown in Figure 16. The model ensemble shows thick volcanic ash plumes close to the Eyjafjallajökull with $m'_{\text{col}} \geq 10 \text{ g m}^{-2}$, and ash clouds with $m'_{\text{col}} < 1 \text{ g m}^{-2}$ extending over large parts of Europe. VADUGS detects only very limited ash clouds, which correspond spatially to the thickest parts of the simulated ash clouds but with a lower m_{col} . However, in some cases it does not detect ash directly at the volcano, although a prominent plume is simulated. VACOS detects more ash contamination, especially also the plumes in the direct surroundings of the vent, and retrieves higher m_{col} for them compared to VADUGS. Still, the model ensemble often produces larger ash-loaded areas in the surroundings of the volcano than retrieved by VACOS. At greater distances VADUGS hardly retrieves any ash, whereas VACOS regularly finds ash clouds above central Europe. For instance, in Figure 16b3 volcanic ash is found above the Atlantic north of Scotland, south-east England and at the Atlantic coast of France; then, the ash cloud bends towards north-east and continues over central France and central Germany with concentrations around $0.2\text{--}0.3 \text{ g m}^{-2}$. The model ensemble produces a similar distribution, but the ash cloud heads further south after the bending, continuing above the Mediterranean sea (panel a3). Similar offsets between the model and the VACOS retrieval are also visible comparing the other two scenes; in both cases there are faint ash clouds that are detected further east by VACOS than they are simulated by the model ensemble. This indicates that the model is less reliable at large distances, e.g., above central Europe, which might also influence the following comparisons, but at the same time it confirms that the satellite retrievals over here are plausible and reliable.

Figure 17 shows the POD and the FAR for the different retrievals and accumulation rules with respect to different subsets of the model ensemble. Furthermore, offsets up to 100 km are taken into account, i.e., an ash-loaded grid cell of the simulation is considered detected if the retrievals find ash within the given radius, and a retrieved ash-loaded cell is considered a false alarm if no ash is modeled within the same radius. VADUGS is found to have the smallest POD and FAR, whereas both metrics are larger for VACOS but decrease for more conservative accumulation and threshold rules. Increasing the offset distance leads to significant increases in POD, whereas the decrease of FAR is mostly rather small. When increasing $m'_{\text{col, thrs}}$, the POD as well as the FAR increase. The latter is explained by samples with $m'_{\text{col}} < m'_{\text{col, thrs}}$ but $m_{\text{col}} > m_{\text{col, thrs}}$, existing either because of overestimation by the retrievals, or in the case of the subset with $m'_{\text{col, thrs}} = 0.2 \text{ g m}^{-2}$ due to $m_{\text{col, thrs}} = 0.1 \text{ g m}^{-2}$, such that samples with correctly retrieved mass loads in between those two values lead to false alarms. Considering $m'_{\text{col}} \geq 0.01 \text{ g m}^{-2}$ and an offset of 100 km leads for VADUGS to a POD of ~30% and a FAR of 0.02%, whereas VACOS (with $m_{\text{col, thrs}} = 0.2 \text{ g m}^{-2}$, $f_{\text{thrs}} = 0.5$) has a POD of close to 60% at a FAR of about 2%. The PODs increase by ~20% for $m'_{\text{col}} \geq 0.1 \text{ g m}^{-2}$ and by 25–30% for $m'_{\text{col}} \geq 0.2 \text{ g m}^{-2}$. In the latter case, the POD of VACOS is nearly 90% and the FAR is 2–3%.

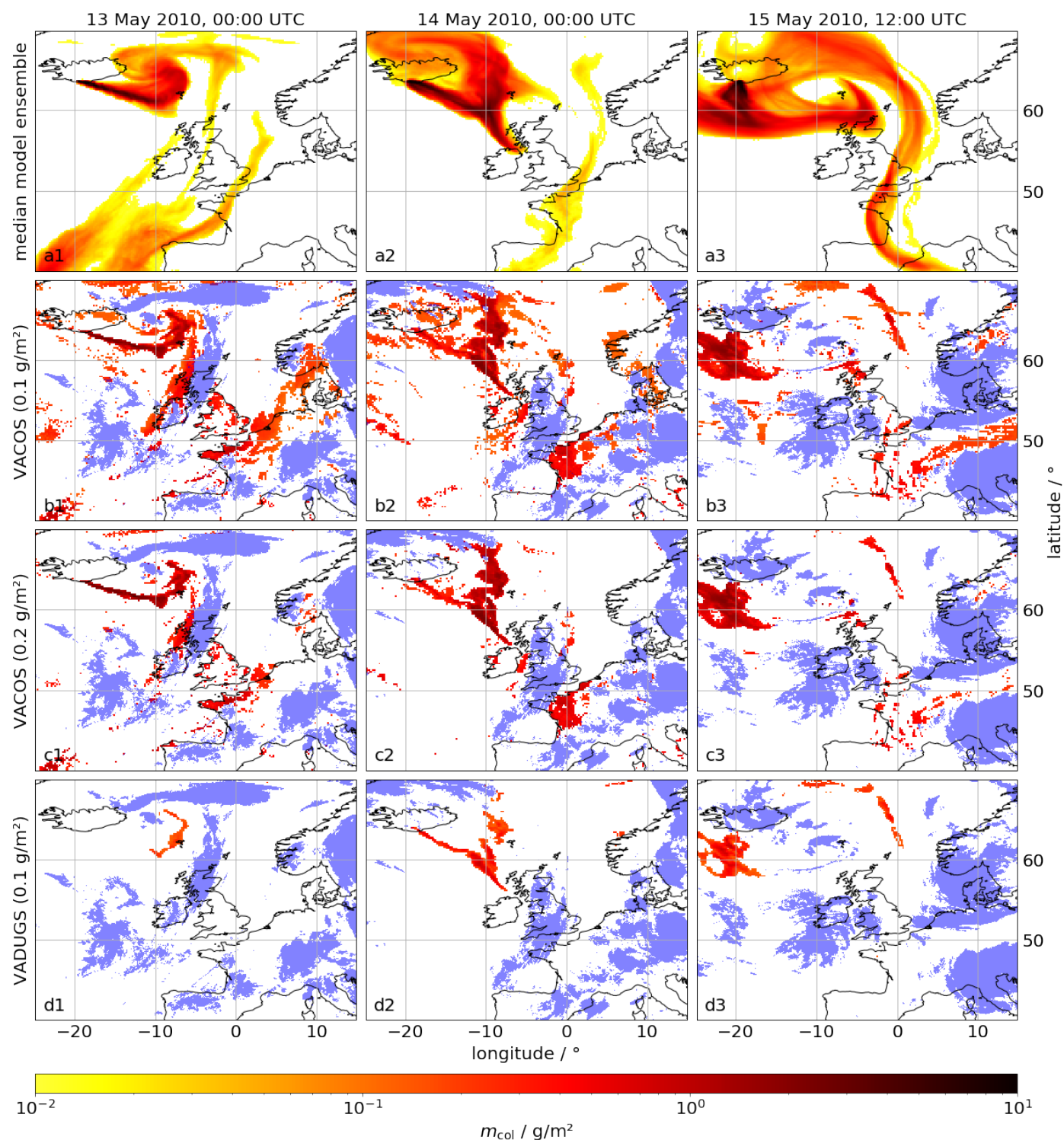


Figure 16. m_{col} (red) according to the median of a model ensemble (a), VACOS ($m_{\text{col,thr}} = 0.1 \text{ g m}^{-2}$ and 0.2 g m^{-2} , $f_{\text{thr}} = 0.5$; b and c) and VADUGS ($m_{\text{col,thr}} = 0.1 \text{ g m}^{-2}$, $f_{\text{thr}} = 0.5$; d) for three scenes (1: 13 May 2010, 00:00 UTC; 2: 14 May 2010, 00:00 UTC; 3: 15 May 2010, 12:00 UTC); for the model $m_{\text{col}} < 0.01 \text{ g m}^{-2}$ is not shown; cirrus presence according to COCS is indicated (blue).

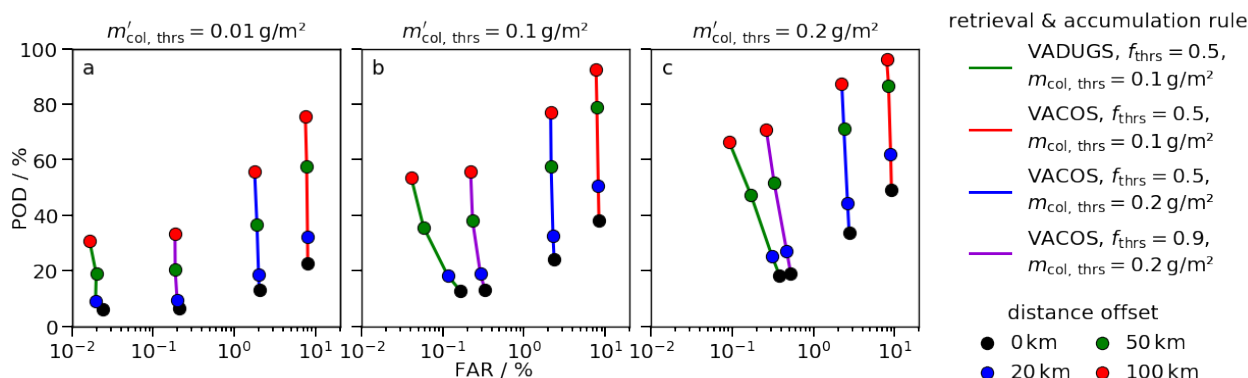


Figure 17. Estimations of the POD and the FAR for the detection of volcanic ash using m'_{col} of VACOS (connected by red, blue and violet lines) and of VADUGS (green line) applying different thresholds and accumulation rules (see text); different thresholds for the model data are used, e.g., only samples with $m'_{col} \geq 0.01 \text{ g m}^{-2}$ (a), $\geq 0.1 \text{ g m}^{-2}$ (b) or $\geq 0.2 \text{ g m}^{-2}$ (c) are classified as ash-containing; only pixels in absence of cirrus clouds are used, ca. 66,000 in total; different offset distances between the model and the satellite data are considered and indicated by the marker color (see text).

Table 5 gives the MAPE and MPE for the retrieval of m_{col} for VADUGS and VACOS, different accumulation and threshold rules and different subsets of the modeled data, not considering any distance offsets. For $m'_{col} \geq 0.01 \text{ g m}^{-2}$ the MAPE is around 100% and the MPE is -96% for VADUGS, whereas for VACOS the MAPE is higher (106–138%) and the MPE is less negative (-26 to -83%), i.e., the underestimation is smaller compared to VADUGS. However, considering only grid cells that are classified as ash-contaminated by the model and the retrieval, i.e., m'_{col} and $m_{col} \geq 0.2 \text{ g m}^{-2}$, the MAPE and MPE of VADUGS change to 65% and -60% , respectively. Using the same accumulation and threshold ($m_{col, thrs} = 0.1 \text{ g m}^{-2}$, $f_{thrs} = 0.5$) for VACOS leads to a MAPE of only 57% and a MPE of 0%; thus, the retrievals of VACOS are in better agreement with the model results than those of VADUGS, indicating that VACOS is more reliable for these mass concentrations.

Table 5. Comparison of the m_{col} retrieval of VADUGS and VACOS against the model ensemble median; MAPE and MPE of different subsets depending on m'_{col} from the model ensemble and m_{col} from the satellite retrieval are considered; for VACOS a mass extinction coefficient at $10.8 \mu\text{m}$ of $200 \text{ m}^2 \text{ kg}^{-1}$ is considered for the conversion of $\tau_{10.8}$ to m_{col} ; different accumulation rules and thresholds are used for the satellite retrievals; pixels with cirrus cloud presence are excluded.

Algorithm	Accumulation Rule ($m_{col, thrs}; f_{thrs}$)	$m'_{col} \geq 0.01 \text{ g m}^{-2}$			$m'_{col} \geq 0.2 \text{ g m}^{-2}$			m'_{col} and $m_{col} \geq 0.2 \text{ g m}^{-2}$		
		Samples	MAPE	MPE	Samples	MAPE	MPE	samples	MAPE	MPE
VADUGS	$0.1 \text{ g m}^{-2}; 0.5$	222,932	99%	-96%	63,595	94%	-94%	6201	65%	-60%
VACOS	$0.1 \text{ g m}^{-2}; 0.5$	221,663	138%	-26%	63,363	79%	-56%	26,200	57%	0%
VACOS	$0.2 \text{ g m}^{-2}; 0.5$	221,633	127%	-52%	63,357	87%	-62%	21,331	60%	12%
VACOS	$0.2 \text{ g m}^{-2}; 0.9$	221,663	106%	-83%	63,363	94%	-76%	12,006	68%	25%

6. Unraveling the Black Box: How Do the ANNs Work?

In this section, we aim to shed light upon the working principles of our ANNs. Therefore, we first perform additional radiative transfer calculations for different ash cloud settings to investigate how the brightness temperatures (differences) depend on the ash cloud properties and show examples on how to deduce those properties from different combinations of channels. Second, we analyze the importance of the different input features of the ANNs with respect to their performance and connect these results with the conclusions drawn from the simulations.

A single, homogeneous ash layer of geometrical thickness 1 km without meteorological clouds is assumed; m_{col} is varied to be $0.1\text{--}1000 \text{ g m}^{-2}$ and z_{top} is 3–12 km above

ground. The volcanic ash has the refractive index of Eyjafjallajökull ash [14], a log-normal size distribution with $r_{\text{eff}} = 0.6 \mu\text{m}$, $3 \mu\text{m}$ and $6 \mu\text{m}$, $s = 1.5$ and a representative shape distribution of spheroids as considered by Piontek et al. [3]. The brightness temperatures of the infrared channels of SEVIRI are simulated for the different ash clouds 500 times for different atmospheric settings and geographical locations in 2010 with the methods outlined in Piontek et al. [2]; finally, the median is determined. Figure 18 shows the dependence of the brightness temperatures on m_{col} . Panel (a) shows (for $z_{\text{top}} = 9 \text{ km}$ and $r_{\text{eff}} = 0.6 \mu\text{m}$) that the brightness temperature decreases with increasing m_{col} . Clearly visible are the atmospheric window channels ($BT_{8.7}$, $BT_{10.8}$, BT_{12}), which exhibit a S-curve behavior with plateaus for $m_{\text{col}} < 0.1 \text{ g m}^{-2}$ and $m_{\text{col}} > 100 \text{ g m}^{-2}$. The water vapor channels ($BT_{6.2}$, $BT_{7.3}$) reach saturation at higher m_{col} . The absolute changes in brightness temperature are largest for the atmospheric window channels, about half as large for $BT_{7.3}$, $BT_{9.7}$, $BT_{13.4}$ and smallest for $BT_{6.2}$. The reason is that all brightness temperatures are around 240 K for a thick ash layer (e.g., $m_{\text{col}} = 100 \text{ g m}^{-2}$) which dominates over all other atmospheric constituents, whereas in the absence of an ash layer the brightness temperature depends on the impact of the atmospheric gases on the different channels. Panel (b) shows the results for $BT_{10.8}$ for $r_{\text{eff}} \in \{0.6, 3, 6\} \mu\text{m}$ and $z_{\text{top}} \in \{3, 6, 9, 12\} \text{ km}$. It shows that with increasing z_{top} the asymptotic value of $BT_{10.8}$ at large m_{col} decreases (from roughly 280 K to 220 K), as the opaque ash layer completely hides the surface and the top of atmosphere brightness temperature is mostly determined by the atmospheric temperature at z_{top} , which decreases with increasing z_{top} within the troposphere. The influence of r_{eff} is smaller and mainly visible in the intermediate regime (i.e., for $m_{\text{col}} \in [1, 30] \text{ g m}^{-2}$), where $BT_{10.8}$ is smallest for $r_{\text{eff}} = 3 \mu\text{m}$ and largest for $6 \mu\text{m}$. This can be understood from the size dependence of the mass extinction coefficient, which is largest for $r_{\text{eff}} = 3 \mu\text{m}$ and smallest for $6 \mu\text{m}$ with respect to the three considered r_{eff} [2]. Note that both an increase in z_{top} and in m_{col} leads to a lower $BT_{10.8}$.

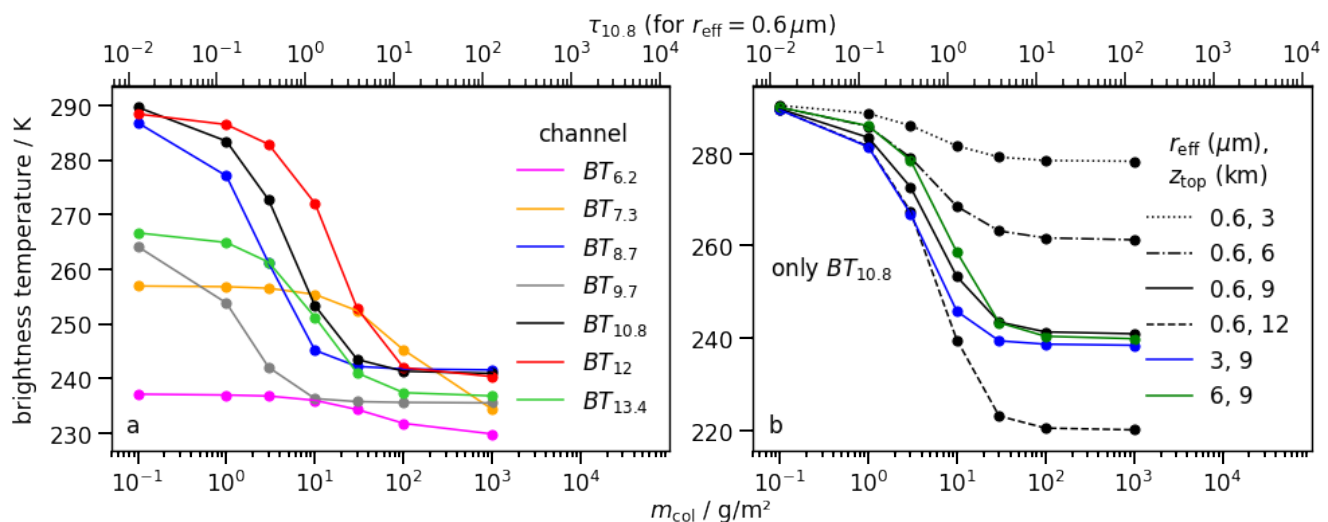


Figure 18. Median brightness temperature of different SEVIRI channels averaged for 500 simulations (see text); (a) brightness temperatures for an ash cloud with $z_{\text{top}} = 9 \text{ km}$ and $r_{\text{eff}} = 0.6 \mu\text{m}$ for different m_{col} , (b) $BT_{10.8}$ for different z_{top} and r_{eff} .

Figure 19 shows combinations of different brightness temperatures and brightness temperature differences for the same simulations. The size of the markers encodes m_{col} . Markers of constant z_{top} and r_{eff} but different m_{col} are connected by black, blue and violet lines for $r_{\text{eff}} = 0.6 \mu\text{m}$, $3 \mu\text{m}$ and $6 \mu\text{m}$, respectively, whereas the linestyle denotes z_{top} . Additionally, points of constant m_{col} , r_{eff} and variable z_{top} are connected by red lines, and points of constant m_{col} , z_{top} and variable r_{eff} by green lines. Thus, Figure 19 shows whether the variation of different parameters leads to similar behaviors with respect to certain brightness temperatures (differences). Panel (a) shows $BT_{10.8}$ against $BTD_{10.8-12}$;

this combination has often been studied [37,38]. For $m_{\text{col}} < 3 \text{ g m}^{-2}$ an increase in m_{col} (along the black lines) reduces $BT_{10.8-12}$ and $BT_{10.8-12}$; the latter is related to the spectral dependence of the mass extinction coefficient of volcanic ash [3]. An increase in z_{top} (along the red line) also reduces $BT_{10.8}$ and leads overall to a similar change in this two-dimensional phase space, i.e., the black and the red curves are parallel and lie on top of each other. However, for $m_{\text{col}} > 3 \text{ g m}^{-2}$ the directions of these curves deviate as $BT_{10.8-12}$ vanishes if the ash cloud becomes opaque in the thermal infrared; thus, m_{col} and z_{top} can be distinguished. In (c), $BT_{13.4}$ and $BT_{12-13.4}$ are combined. Here the black and the red curve proceed in different directions already for $m_{\text{col}} < 1 \text{ g m}^{-2}$ and allow the separation of the influence of m_{col} and z_{top} : Increasing either of the two quantities reduces $BT_{13.4}$, but increasing z_{top} changes BT_{12} and $BT_{13.4}$ similarly, thus $BT_{12-13.4}$ shows only minor changes. Yet, increasing m_{col} leads to a reduction of $BT_{12-13.4}$, which is $\sim 20 \text{ K}$ in the absence of volcanic ash due to the impact of carbon dioxide (see Figure 18), but vanishes as the ash cloud becomes opaque. In principle, the same behavior is visible for $r_{\text{eff}} = 3 \mu\text{m}$ in (b) and (d), but the curves corresponding to different z_{top} are located closer together. Note that for $r_{\text{eff}} = 6 \mu\text{m}$ (violet curve) $BT_{10.8-12}$ is positive independent of m_{col} for $z_{\text{top}} = 9 \text{ km}$ [3,38].

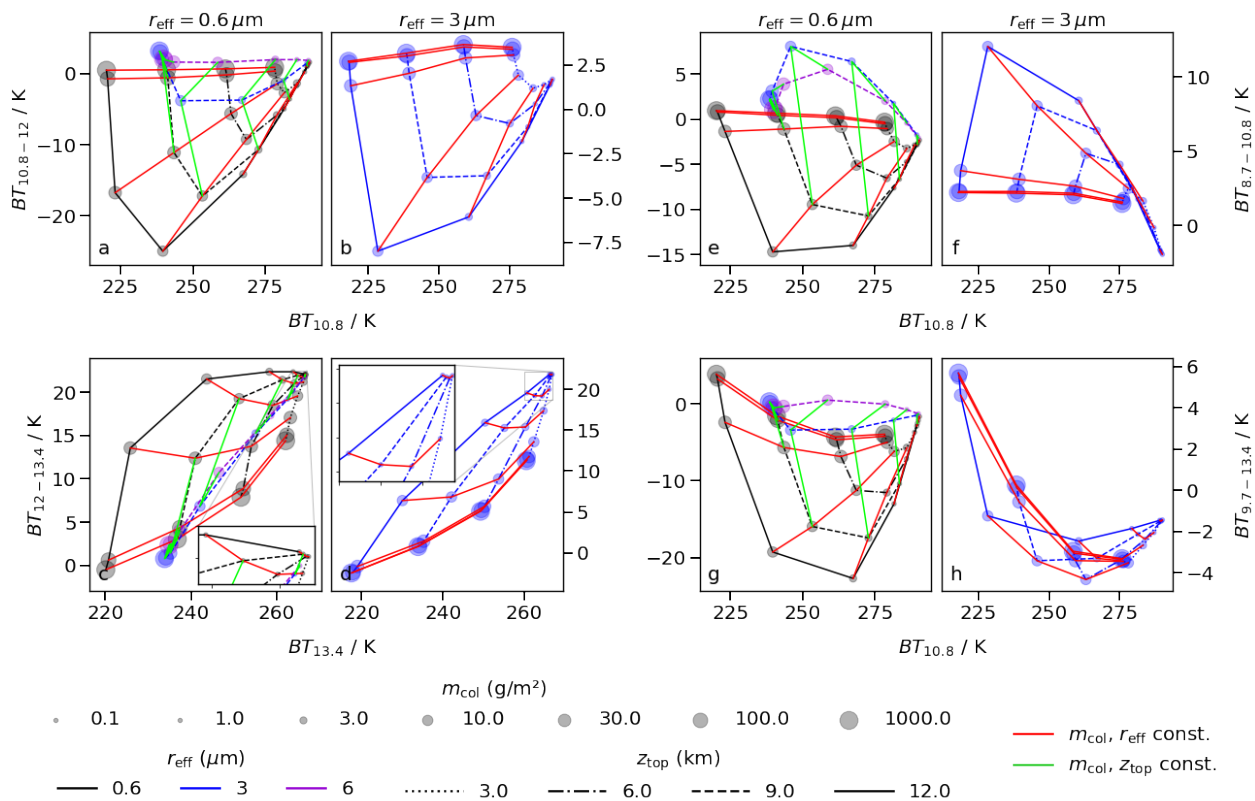


Figure 19. Combinations of the medians of different brightness temperatures and brightness temperature differences, calculated from 500 simulations for each volcanic ash cloud (see text); mainly two different r_{eff} are used: $0.6 \mu\text{m}$ (a,c,e,g) and $3 \mu\text{m}$ (b,d,f,h); m_{col} is given by the size of the markers; markers with constant z_{top} and r_{eff} but different m_{col} are connected by black, blue and violet lines for $r_{\text{eff}} = 0.6 \mu\text{m}$, $3 \mu\text{m}$ and $6 \mu\text{m}$, respectively; the linestyle encodes z_{top} ; points of constant m_{col} , r_{eff} and variable z_{top} are connected by red lines, and points of constant m_{col} , z_{top} and variable r_{eff} by green lines.

The variation of brightness temperatures (differences) due to r_{eff} is shown by the green lines as an example. In (a) and (c), r_{eff} cannot be determined easily as it is entangled with m_{col} and z_{top} . Therefore, we consider $BT_{8.7}$, $BT_{9.7}$ and $BT_{10.8}$, which are located at the typical absorption peak of volcanic ash and which are influenced differently depending on the particle size [3]. Panel (e) shows $BT_{10.8}$ against $BT_{8.7-10.8}$. Here $r_{\text{eff}} = 0.6 \mu\text{m}$ mostly leads to $BT_{8.7-10.8} < 0$, in contrast to $r_{\text{eff}} = 3 \mu\text{m}$ and $6 \mu\text{m}$ (except for $m_{\text{col}} < 1 \text{ g m}^{-2}$, but even

there the r_{eff} can be separated). Panel (g) shows $BT_{D_{9.7-13.4}}$ on the y-axis. The blue curve for $r_{\text{eff}} = 3 \mu\text{m}$ is convex, whereas the violet curve for $r_{\text{eff}} = 6 \mu\text{m}$ is concave. A threshold at roughly -1.5 K with respect to $BT_{D_{9.7-13.4}}$ would allow to separate the different r_{eff} except for $m_{\text{col}} > 30 \text{ g m}^{-2}$. To summarize, it is possible to disentangle the physical quantities m_{col} , z_{top} and r_{eff} using the MSG/SEVIRI channels in the thermal infrared and exploiting the dependence of the mass extinction coefficient of volcanic ash on the wavelength and the particle size as well as the relatively constant impact of other atmospheric gases (e.g., carbon dioxide).

The actual working principles of ANNs are hard to determine. However, one can try to quantify the importance of the individual input features for each ANN and thereby deduce which physical principles are exploited and which functions might be implemented internally. Neglecting those features and retraining the ANNs might also allow to simplify the algorithms [39,40]. We define two metrics: The first, M_x , is the relative contribution of the x th input neuron to the total weight between the input layer and the first hidden layer, defined as

$$M_x = \frac{\sqrt{\sum_{j=0}^m w_{x,j}^2}}{\sum_{i=0}^n \sqrt{\sum_{j=0}^m w_{i,j}^2}} \quad (1)$$

for an input layer of n neurons, a first hidden layer of m neurons and the connecting weights $w_{i,j}$ [10]. The expectation is that the weights of unimportant features will vanish during training such that they have no impact on the calculation of the ANNs. However, in case of multiple hidden layers it might be possible that the impact of a feature significantly changes in the subsequent layers. Therefore, we consider a second metric: the relative change in loss, K_x , when a feature is set to zero (simulating $w_{x,j} = 0$ for all $j \in [1, m]$) for the complete test data set, i.e.,

$$K_x = \frac{L_x - L_0}{L_0} \quad (2)$$

with L_0 being the loss (here the mean squared error for regressions and the categorical cross entropy for the classification) for the full test data set, whereas L_x is the loss for the test data set when setting the x th input to zero. When dropping an unimportant feature L_x should not change significantly, no matter whether it has vanishing weights already between the input and the first hidden layer or deeper in the network. However, dropping an important feature that is necessary for the calculation of the ANN will lead to a worse performance and, therefore, a larger loss L_x .

Figure 20 shows M_x and K_x for all four ANNs and their input features. The two metrics quantitatively lead to different results. Whereas M_x shows for all features always at least a small value, K_x produces relatively large contrasts; thus, using K_x it becomes more obvious whether an input is important or not. However, qualitatively both methods lead to similar pictures: For example, for the classification both metrics (a,b) indicate $BT_{10.8}$ to be the most important of all brightness temperatures. Compared with the results of the other ANNs also the total column water and the total column water vapor appear to be important, and finally the viewing zenith angle.

For the $\tau_{10.8}$ retrieval (c,d) only ash-loaded samples are considered, with BT_{12} being the most important channel. Furthermore, $BT_{8.7}$, $BT_{10.8}$ and $BT_{13.4}$ are prominent in both metrics. This is in agreement with the conclusions drawn from Figure 19. Compared to the retrievals of z_{top} and r_{eff} also the total column water and water vapor have significant impact on K_x (d); they might be used to take the corresponding atmospheric constituents into account. For the z_{top} retrieval, M_x (e) implies $BT_{8.7}$, $BT_{9.7}$, $BT_{10.8}$ and $BT_{13.4}$ to be the most important brightness temperatures, and furthermore $\tau_{10.8}$ and surprisingly the day of year variables have large metrics. However, generally the contrast is not very large between the features. K_x (f) however points out $BT_{9.7}$, $BT_{13.4}$ and $\tau_{10.8}$. The importance of $BT_{13.4}$ is again in agreement with the conclusions from Figure 19. The large values for $\tau_{10.8}$ show that the performance of the z_{top} algorithm heavily depends on the accuracy of the $\tau_{10.8}$ retrieval. For the r_{eff} retrieval, M_x (g) slightly highlights $BT_{8.7}$, BT_{12} , $BT_{13.4}$.

From the auxiliary features the latitude, the cosine of the viewing zenith angle, $\tau_{10.8}$ and the ash-free temperatures $BT_{8.7, \text{clr}}$, $BT_{10.8, \text{clr}}$ and $BT_{12, \text{clr}}$ stand out. Again, the contrast between features is overall small. Using K_x (h) leads roughly to the same result but stresses the leading channels $BT_{8.7}$, BT_{12} , $BT_{13.4}$ even more. The importance of these channels for the derivation of r_{eff} was also visible in Figure 19.

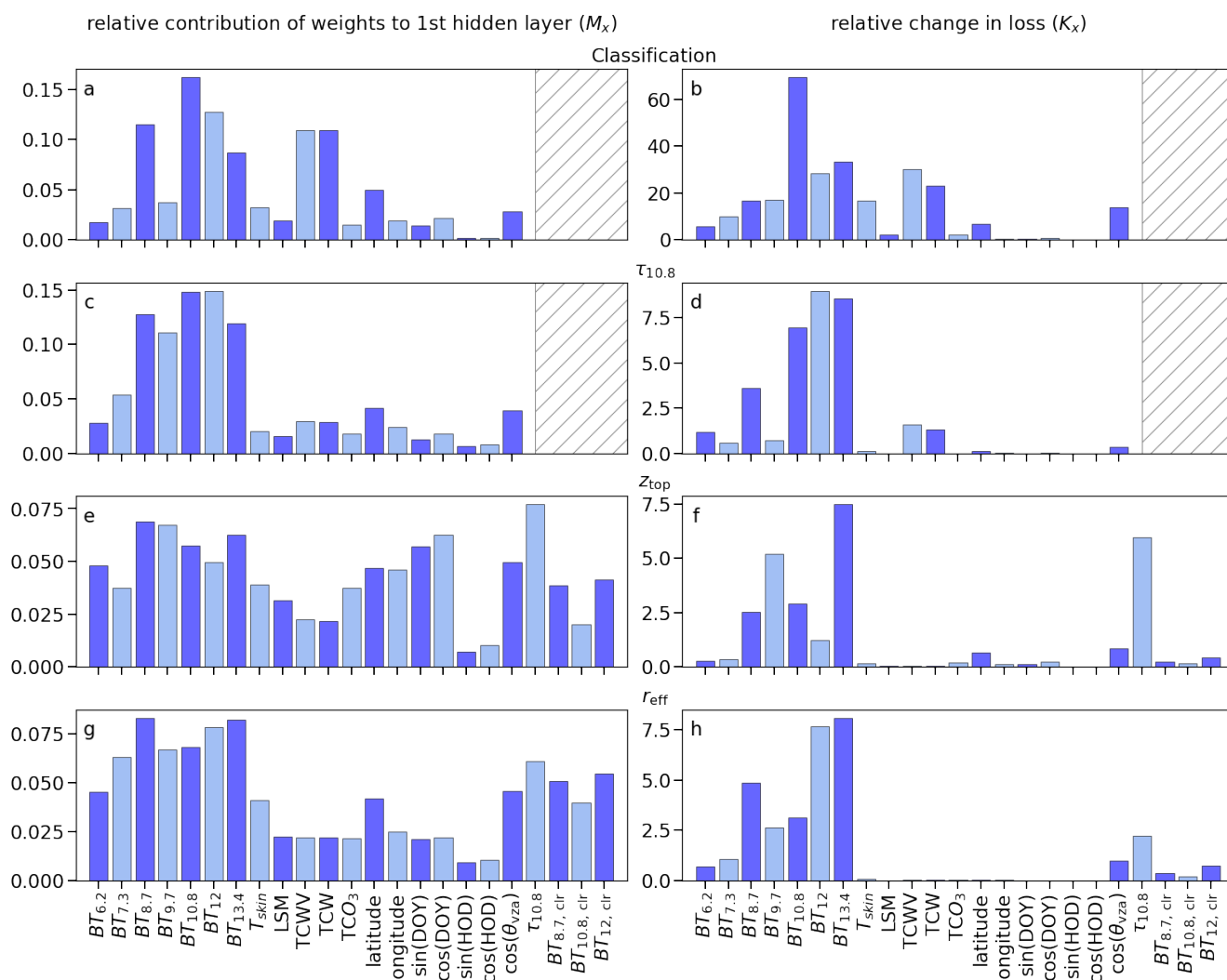


Figure 20. Two estimations of the feature importance: the relative contribution to the total weight at the 1st hidden layer (M_x ; **a,c,e,g**) and the relative change in loss (K_x ; **b,d,f,h**) (see text for definition) for all four ANNs and all input features, including the viewing zenith angle (θ_{vza}), the hour of day (HOD) and the day of year (DOY); the loss function used for K_x is the mean squared error for regressions and the categorical cross entropy for classification; bars are alternately colored blue and light blue for better readability.

Let us focus on K_x (b, d, f, h) now: From the auxiliary input features the total column water vapor and the total column water play a minor role in the retrieval of $\tau_{10.8}$ and a larger one for the classification; for height and effective radius retrievals they are unimportant. The latitude appears to be important mostly for the z_{top} retrieval and the classification, and the skin temperature only for the classification. The land/sea mask, the total column ozone, the longitude, the day of year and the hour of day are rather negligible in all cases.

The metric M_x was also derived for the input features of the cirrus cloud retrieval CiPS, which also consisted of several ANNs to derive, e.g., the cirrus optical depth or the cloud top height, trained using collocated SEVIRI and CALIOP measurements [10]. Comparing the results CiPS and VACOS shows similarities, e.g., in both cases the brightness temperatures play a dominant role, but the day of year and surface classifications are rather

negligible. A noteworthy difference is that whereas the z_{top} retrieval of CiPS mainly depends on latitude (as maximum cirrus cloud top height strongly depends thereon; thus, a statistical effect), our results show only a smaller dependence on this quantity.

Generally, the metrics support the observations made before in Figure 19, with the brightness temperatures having the major impact, whereas the auxiliary data seem to be of minor importance (even the skin temperature). The low values for the geographical coordinates imply that the ANNs do not learn the geography of the Earth as visible from SEVIRI. Similarly, the low values for the land/sea mask show that even this rough classification of the Earth's surface remains unregarded. The low values for the times (day of year, hour of day) indicate that the ANNs do also not learn seasonal or diurnal variations. Reasons why the ANNs do not internalize those more evolved concepts might be a too small training data set, or they are just not helpful enough as the central physical quantities are rather obtained in the observations. For example, although the atmospheric state undergoes a seasonal variation, volcanic ash clouds are independent of them as volcanic eruptions can take place at any time of year. The hint that the ANNs do not learn the map indicates that the method might be applicable to other regions of the Earth (e.g., using GOES or Himawari satellites).

7. Conclusions

In a companion paper, we introduced a new algorithm to retrieve volcanic ash properties, i.e., a pixel classification, the cloud top height (z_{top}), the effective particle radius (r_{eff}) and (indirectly from the optical depth at $10.8\ \mu\text{m}$, $\tau_{10.8}$) the mass column concentration (m_{col}) from MSG/SEVIRI data using artificial neural networks; it is called VACOS (*Volcanic Ash Cloud properties Obtained from SEVIRI* [2]). The input data encompass the seven brightness temperatures of the imager's channels in the thermal infrared and additional data from ECMWF. VACOS allows spatially and temporally highly resolved retrievals of volcanic ash clouds independent of daylight.

For the validation, VACOS is compared to independent measurements. With respect to CALIOP retrievals of the Puyehue-Cordón Caulle ash clouds (2011), VACOS shows quite a large variability on a single pixel basis. A regional average over 5×5 pixels reduces this variability and leads to lower deviations between m_{col} of VACOS and the lidar retrieval: the mean absolute percentage error (MAPE) of m_{col} decreases to 45% and the MAPE of z_{top} is 18% when considering only samples with a lidar retrieval of $m_{\text{col}} \geq 0.2\ \text{g m}^{-2}$. Therefore, this averaging is recommended for future applications of the new algorithm. The VACOS-derived m_{col} is in good agreement with the CALIOP measurements in most, but not all, cases. Deviations are generally within the uncertainties of the reference data. z_{top} and r_{eff} by VACOS have the correct order of magnitude but exhibit significant scattering, with z_{top} being slightly underestimated (having a mean percentage error of -14%), whereas r_{eff} lies within the regime constrained by literature values. The four category classification of VACOS detects the volcanic ash clouds with the highest mass loads, but ash layers with m_{col} less than $\sim 0.4\ \text{g m}^{-2}$ are often misclassified as meteorological clouds (which includes liquid and ice water clouds). Similarly, SEVIRI images indicate that clear sky might be too often misclassified as cloudy, but here dedicated retrievals of meteorological clouds should provide help. Additionally considering a cirrus retrieval also shows that thick ice clouds have the potential to completely hide the volcanic ash or distort their retrieval; thus, we also recommend to consider VACOS results always together with a cirrus cloud retrieval to avoid misinterpretations.

As another reference, airborne lidar and in situ data of an Eyjafjallajökull ash cloud (2010) are used as obtained during two different measurement campaigns by FAAM and DLR, respectively. The various measurement results for m_{col} , z_{top} and r_{eff} agree well, considering that different instrumentation was used and different parts of the ash cloud were probed. Main differences are a higher upper estimate of m_{col} according to the DLR (roughly $2\ \text{g m}^{-2}$ compared to mostly around $1\ \text{g m}^{-2}$) and a higher upper limit for r_{eff} in the FAAM data ($2\ \mu\text{m}$ compared to $1\ \mu\text{m}$); z_{top} estimates are up to $\sim 6\ \text{km}$ in both cases.

For the most part, the VACOS retrievals of m_{col} and r_{eff} lie within the uncertainty intervals of the airborne measurements, whereas z_{top} from the new retrieval scatters around the reference values with deviations in the order of 1–2 km. Concurrent overestimations of z_{top} and r_{eff} can be explained by low m_{col} in some cases, but not generally. The classification algorithm correctly detects the ash cloud.

Results of a volcanic ash transport and dispersion multi model multi source term ensemble simulating the eruption of Eyjafjallajökull (2010) are used to analyze the performance of VACOS on large scales. Different accumulation rules are investigated when regridding the satellite retrievals. Example scenes indicate that the model results and the VACOS retrievals are in good agreement close to the volcano, but that there are displacements of the ash clouds at larger distances, e.g., for ash clouds above continental Europe; those might be caused by inaccuracies in the model calculations. As a consequence, distance offsets up to 100 km are considered. The model ensemble covers a larger interval of possible m_{col} than VACOS: m_{col} can be an order of magnitude larger in the model calculations than in the VACOS retrieval in a surrounding of some 100 km around the vent, whereas it can be more than one order of magnitude smaller at distances of >1000 km. Considering only samples with $m_{\text{col}} \geq 0.2 \text{ g m}^{-2}$ according to both the model ensemble and VACOS results in a MAPE of ~60% for m_{col} of VACOS.

To further quantify the performance of VACOS, a simulated test data set similar to its training data is used. The four category classification classifies correctly more than 94% of the simulated cases, except when both volcanic ash and meteorological clouds are present, which reduces the amount of correct classifications to ~50%. Simplifying classification results to a binary ash flag results in a probability of detection (POD) of close to 100%. The retrievals of $\tau_{10.8}$, z_{top} and r_{eff} have mostly MAPEs of 10–100%. For $\tau_{10.8}$, the MAPE is ~40% or less for ash layers with a true $\tau_{10.8}$ of 0.1 (corresponding to m_{col} of $0.3\text{--}0.7 \text{ g m}^{-2}$) or more; the retrieval error of z_{top} is up to ~10% for ash layers above 5 km; and r_{eff} has an error of up to 35% for true radii of $0.6\text{--}6 \mu\text{m}$. The performance increases with increasing $\tau_{10.8}$ and z_{top} of the ash layers. Thus, the greatest errors occur for the thinnest and lowest ash clouds, for which the MAPE can even exceed 100%. No significant differences exist between underlying land and sea surfaces, in contrast to the presence of meteorological clouds (particularly if they are located above the ash layer), which can increase the MAPE by up to one order of magnitude and decrease the POD by a factor of two. Analysis of the geographical dependence shows that deserts lead to a decreased performance for clear sky cases, e.g., the accuracy of the binary ash flag slightly decreases (ca. 1%), whereas the MAPE for r_{eff} increases. An increased performance of the binary ash flag is observable in the presence of clouds above areas that are typically cloudy, i.e., where many cloudy samples are included in the simulated (training and test) data sets. The retrievals of $\tau_{10.8}$ and r_{eff} show increased MAPEs with increasing viewing zenith angle, whereas the MAPE of z_{top} increases with the absolute latitude. Using further simulated test data sets, the dependence of the retrievals with respect to the ash layer(s) is investigated. The (unretrieved) geometrical layer thickness and the presence of multiple layers might introduce errors of about 30% for $\tau_{10.8}$, 5% for z_{top} , but are negligible for r_{eff} .

Volcanic ash detection can be performed using the binary ash flag or $\tau_{10.8}/m_{\text{col}}$ and corresponding thresholds. Using the simulated data shows that the detection ability of the binary ash flag is better than the one using $\tau_{10.8}$, and both outperform the usage of the brightness temperature difference between the channels centered at $10.8 \mu\text{m}$ and $12 \mu\text{m}$; using for m_{col} a threshold of 0.2 g m^{-2} and considering only ash layers with $0.2\text{--}1 \text{ g m}^{-2}$ leads to a POD of more than 90% and a false alarm rate (FAR) of ca. 1%. For the CALIOP data, higher PODs can be found when using m_{col} compared to the binary ash flag; a similar threshold as before leads to a POD of ~70%. Using the model results as reference data, we find for similar thresholds PODs of 20–85% and FARs of 0.3–3%, which strongly depend on the accumulation rule and the allowed distance offsets. Note that the model result is not a perfect representation of reality; therefore, POD and FAR should be regarded as metrics

quantifying the agreement of model and retrieval results rather than performance metrics attributed exclusively to VACOS.

VACOS is also compared to its direct predecessor VADUGS. The latter exhibits a better detection performance on the CALIOP data and has a smaller FAR for the model data, but it also misses significantly more modeled ash clouds than VACOS, especially also ash plumes close to the vent. Whereas the m_{col} retrieval of VADUGS performs similar to VACOS for the CALIOP data, it underestimates m_{col} compared to the airborne lidar and in situ findings. The VADUGS retrieval of z_{top} is reliable only for ash clouds with m_{col} of the order 1 g m^{-2} or larger. Overall, VADUGS seems much more sensitive to specific cases than VACOS. In the end, we consider the importance of different input features of the VACOS retrievals, showing that they are mostly dependent on the SEVIRI brightness temperatures, and partly also on the viewing zenith angle as well as total column water and water vapor estimates from ECMWF. Longitude, land/sea-mask and times have negligible impact on the output.

In the future it would be desirable to further analyze the retrieval performance also with respect to mineral dust. As volcanic ash and mineral dust share similar optical properties due to the common high silica content, it is likely that VACOS might misclassify dust as volcanic ash [7,41,42]. In this case, it would be interesting to see if VACOS could be used to retrieve dust cloud properties as well. The aviation industry considers similar risks for volcanic ash and for dust, hence a combined retrieval seems advantageous. Furthermore, VACOS has been tailored for Meteosat-9/MSG2. The usage of the retrieval algorithm with the currently operational MSG satellite as well as the satellites in other operation modes (i.e., rapid scan mode, Indian Ocean data coverage) should be investigated. As other infrared imagers aboard geostationary weather satellites such as GOES-R [43], Himawari-8/9 [44] and Fengyun-4A [45] share similar channels with MSG/SEVIRI one could also investigate the transferability of VACOS to those instruments. The error of the VACOS retrievals decreases when a local average is calculated. To expand on this idea, further processing of the retrieval map seems desirable, e.g., to cluster ash-containing pixels and quantify the resulting ash patches, or even track them in time. Possible fields of application of VACOS include the Volcanic Ash Advisory Centers, the intercomparison with other volcanic ash retrievals (as in [46]), calibrating and validating volcanic ash transport and dispersion models [35,47–49] and flight planning for future in situ measurements [15]. Due to the high spatial and temporal resolution, it can be used to track individual ash clouds to investigate their lifecycle on timespans of days to weeks. In combination with information on liquid and ice water clouds, aerosol-cloud interaction could be analyzed.

In summary, VACOS is well characterized and shown to be reliably applicable under different atmospheric conditions and for various kinds of volcanic ash clouds. It can be utilized for atmospheric research as well as for air space monitoring with respect to volcanic ash. Operational use by the German weather service (DWD) as a follow-on of VADUGS is planned.

Author Contributions: Conceptualization, D.P. and L.B.; methodology, D.P.; software, D.P.; validation, D.P.; formal analysis, D.P.; investigation, D.P.; data sources, J.K., U.S., F.M., M.P. and L.B.; writing—original draft preparation, D.P.; writing—review and editing, L.B., C.V., J.K., U.S., F.M. and M.P.; visualization, D.P.; supervision, L.B. and C.V.; project administration, L.B.; funding acquisition, C.V. and L.B. All authors have read and agreed to the published version of the manuscript.

Funding: D. Piontek and L. Bugliaro were supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 723986 (EUNADICS-AV). C. Voigt acknowledges funding by the Helmholtz Association under contract HGF W2/W3-60.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data presented in this study is available on request from the corresponding author.

Acknowledgments: The MSG/SEVIRI data were provided by EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites) and the auxiliary data were obtained from ECMWF (European Center for Medium-Range Weather Forecasts). Radiative transfer calculations have been performed on DLR's CARA (Deutsches Zentrum für Luft- und Raumfahrt; Cluster for Advanced Research in Aerospace). Airborne data were obtained by DLR using the Falcon 20E, operated by DLR flight experiment facilities, and using the BAe-146-301 Atmospheric Research Aircraft which in 2010 was operated by Directflight Ltd. and managed by FAAM (Facility for Airborne Atmospheric Measurements), a joint entity of NERC (Natural Environment Research Council) and the Met Office. Furthermore, we like to thank M. Vázquez-Navarro, J. Strandgren, B. Mayer and M. Rapp for constructive discussions and valuable feedback, W. Beer for technical support, as well as J. Wilzewski and two reviewers for the helpful comments to improve the manuscript. J. Kar acknowledges useful discussions with M. Vaughan on correcting the low bias from the default ash lidar ratios used in CALIOP version 4.10.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Metrics

Different error metrics are used in this work, derived from a set of n pairs of retrieved values r_i and corresponding true values t_i . The *mean squared error* MSE is defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (r_i - t_i)^2. \quad (\text{A1})$$

The *mean absolute percentage error* MAPE is calculated by

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{r_i - t_i}{t_i} \right|. \quad (\text{A2})$$

The *mean percentage error* MPE is

$$\text{MPE} = \frac{100}{n} \sum_{i=1}^n \frac{r_i - t_i}{t_i}. \quad (\text{A3})$$

To quantify the performance of a boolean retrieval we consider the *probability of detection* POD and the *false alarm rate* FAR (also *probability of false detection* [50])

$$\text{POD} = \frac{N_{tp}}{N_{tp} + N_{fn}} \quad (\text{A4})$$

$$\text{FAR} = \frac{N_{fp}}{N_{fp} + N_{tn}} \quad (\text{A5})$$

with N_{tp} being the number of true positives (here meaning that the retrieval signals the presence of volcanic ash, which is really present), N_{fp} the false positives (presence of ash is signaled although none is present), N_{tn} the true negatives (absence of ash is signaled and none is present) and N_{fn} the false negatives (absence of ash is signaled although it is present). The performance of multi-category classifications is described by the *accuracy* given as the number of correctly classified samples divided by the total number of samples. For two categories, the *accuracy* simplifies to

$$\text{accuracy} = \frac{N_{tp} + N_{tn}}{N_{tp} + N_{fp} + N_{tn} + N_{fn}} \quad (\text{A6})$$

and if no negative samples are present, i.e., $N_{tn} = N_{fn} = 0$, the *accuracy* equals the POD.

Appendix B. II-Sigmoid Distribution

The II-sigmoid distribution was introduced by Alivanoglou and Likas [18] and is defined as the difference of two sigmoid functions, i.e.,

$$\Pi(x) = \frac{1}{b-a} \left(\frac{1}{1 + e^{-\lambda(x-a)}} - \frac{1}{1 + e^{-\lambda(x-b)}} \right) \quad (\text{A7})$$

with a and $b > a$ parameterizing the positions of the rise of the corresponding sigmoid, and $\lambda > 0$ describing their steepness. The expectation value is $(a + b)/2$ and for the standard deviation σ holds

$$\sigma^2 = \frac{b^3 - a^3}{3(b-a)} - \frac{a^2 + b^2}{4} - \frac{ab}{2} + \frac{\pi^2}{3\lambda^2}. \quad (\text{A8})$$

References

- Bugliaro, L.; Piontek, D.; Kox, S.; Schmidl, M.; Mayer, B.; Müller, R.; Vázquez-Navarro, M.; Gasteiger, J.; Kar, J. Combining radiative transfer calculations and a neural network for the remote sensing of volcanic ash using MSG/SEVIRI. 2021, in preparation.
- Piontek, D.; Bugliaro, L.; Schmidl, M.; Zhou, D.K.; Voigt, C. The New Volcanic Ash Retrieval VACOS Using MSG/SEVIRI and Artificial Neural Networks: 1. Development. *Remote Sens.* **2021**, *13*, 3112. [CrossRef]
- Piontek, D.; Hornby, A.; Voigt, C.; Bugliaro, L.; Gasteiger, J. Determination of complex refractive indices and optical properties of volcanic ashes in the thermal infrared based on generic petrological compositions. *J. Volcanol. Geotherm. Res.* **2021**, *411*, 107174. [CrossRef]
- Zhou, D.K.; Larar, A.M.; Liu, X.; Smith, W.L.; Strow, L.L.; Yang, P.; Schlüssel, P.; Calbet, X. Global Land Surface Emissivity Retrieved From Satellite Ultraspectral IR Measurements. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 1277–1290. [CrossRef]
- Zhou, D.K.; Larar, A.M.; Liu, X. MetOp-A/IASI Observed Continental Thermal IR Emissivity Variations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 1156–1162. [CrossRef]
- Zhou, D.K.; Larar, A.M.; Liu, X. On the relationship between land surface infrared emissivity and soil moisture. *J. Appl. Remote Sens.* **2018**, *12*, 1–14. [CrossRef]
- Watkin, S.C. The application of AVHRR data for the detection of volcanic ash in a Volcanic Ash Advisory Centre. *Meteorol. Appl.* **2003**, *10*, 301–311. [CrossRef]
- Kaufman, Y.J.; Koren, I.; Remer, L.A.; Rosenfeld, D.; Rudich, Y. The effect of smoke, dust, and pollution aerosol on shallow cloud development over the Atlantic Ocean. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 11207–11212. [CrossRef] [PubMed]
- Strandgren, J.; Bugliaro, L.; Sehnke, F.; Schröder, L. Cirrus cloud retrieval with MSG/SEVIRI using artificial neural networks. *Atmos. Meas. Tech.* **2017**, *10*, 3547–3573. [CrossRef]
- Strandgren, J.; Fricker, J.; Bugliaro, L. Characterisation of the artificial neural network CiPS for cirrus cloud remote sensing with MSG/SEVIRI. *Atmos. Meas. Tech.* **2017**, *10*, 4317–4339. [CrossRef]
- Thomas, G.E.; Stamnes, K. *Radiative Transfer in the Atmosphere and Ocean*, 1st ed.; Cambridge University Press: New York, NY, USA, 2002.
- de Laat, A.; Vázquez-Navarro, M.; Theys, N.; Stammes, P. Analysis of properties of the 19 February 2018 volcanic eruption of Mount Sinabung in S5P/TROPOMI and Himawari-8 satellite data. *Nat. Hazards Earth Syst. Sci.* **2020**, *20*, 1203–1217. [CrossRef]
- Schneider, D.J.; Van Eaton, A.R.; Wallace, K.L. Satellite observations of the 2016–2017 eruption of Bogoslof volcano: Aviation and ash fallout hazard implications from a water-rich eruption. *Bull. Volcanol.* **2020**, *82*, 29. [CrossRef]
- Deguine, A.; Petitprez, D.; Clarisse, L.; Guðmundsson, S.; Outes, V.; Villarosa, G.; Herbin, H. Complex refractive index of volcanic ash aerosol in the infrared, visible, and ultraviolet. *Appl. Opt.* **2020**, *59*, 884–895. [CrossRef]
- Schumann, U.; Weinzierl, B.; Reitebuch, O.; Schlager, H.; Minikin, A.; Forster, C.; Baumann, R.; Sailer, T.; Graf, K.; Mannstein, H.; et al. Airborne observations of the Eyjafjalla volcano ash cloud over Europe during air space closure in April and May 2010. *Atmos. Chem. Phys.* **2011**, *11*, 2245–2279. [CrossRef]
- Marengo, F.; Johnson, B.; Turnbull, K.; Newman, S.; Haywood, J.; Webster, H.; Ricketts, H. Airborne lidar observations of the 2010 Eyjafjallajökull volcanic ash plume. *J. Geophys. Res. Atmos.* **2011**, *116*. [CrossRef]
- Turnbull, K.; Johnson, B.; Marengo, F.; Haywood, J.; Minikin, A.; Weinzierl, B.; Schlager, H.; Schumann, U.; Leadbetter, S.; Woolley, A. A case study of observations of volcanic ash from the Eyjafjallajökull eruption: 1. In situ airborne observations. *J. Geophys. Res. Atmos.* **2012**, *117*. [CrossRef]
- Alivanoglou, A.; Likas, A. Probabilistic Models Based on the II-Sigmoid Distribution. In *Artificial Neural Networks in Pattern Recognition*; Prevost, L., Marinai, S., Schwenker, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 36–43.
- Typical Radiometric Noise, Calibration Bias and Stability for Meteosat-8, -9, -10 and -11 SEVIRI. European Organisation for the Exploitation of Meteorological Satellites. 2019. Available online: https://www-cdn.eumetsat.int/files/2020-04/pdf_typ_radiomet_acc_msg-1-2.pdf (accessed on 5 August 2021).

20. Winker, D.M.; Vaughan, M.A.; Omar, A.; Hu, Y.; Powell, K.A.; Liu, Z.; Hunt, W.H.; Young, S.A. Overview of the CALIPSO Mission and CALIOP Data Processing Algorithms. *J. Atmos. Ocean. Technol.* **2009**, *26*, 2310–2323. [\[CrossRef\]](#)
21. Kim, M.H.; Omar, A.H.; Tackett, J.L.; Vaughan, M.A.; Winker, D.M.; Trepte, C.R.; Hu, Y.; Liu, Z.; Poole, L.R.; Pitts, M.C.; et al. The CALIPSO version 4 automated aerosol classification and lidar ratio selection algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 6107–6135. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Gasteiger, J.; Groß, S.; Freudenthaler, V.; Wiegner, M. Volcanic ash from Iceland over Munich: mass concentration retrieved from ground-based remote sensing measurements. *Atmos. Chem. Phys.* **2011**, *11*, 2209–2223. [\[CrossRef\]](#)
23. Winker, D.M.; Liu, Z.; Omar, A.; Tackett, J.; Fairlie, D. CALIOP observations of the transport of ash from the Eyjafjallajökull volcano in April 2010. *J. Geophys. Res. Atmos.* **2012**, *117*. [\[CrossRef\]](#)
24. Bignami, C.; Corradini, S.; Merucci, L.; de Michele, M.; Raucoules, D.; De Astis, G.; Stramondo, S.; Piedra, J. Multisensor Satellite Monitoring of the 2011 Puyehue-Cordon Caulle Eruption. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2786–2796. [\[CrossRef\]](#)
25. Ishimoto, H.; Hayashi, M.; Mano, Y. Optimal ash particle refractive index model for simulating the brightness temperature spectrum of volcanic ash clouds from satellite infrared sounder measurements. *Atmos. Meas. Tech. Discuss.* **2021**, *2021*, 1–28. [\[CrossRef\]](#)
26. Debling, F.; Schneider, J.F.; Rosi, M.; Leoz-Garziandia, E.; Rorije, E. Technical Cooperation Mission, Effects of the Puyehue-Cordón Caulle Eruption Argentina, 4–19 July 2011. Joint UNEP/OCHA Environment Unit, 2011. Available online: <https://www.eccentre.org/wp-content/uploads/2019/06/Argentina-volcan-eruption-2011-report.pdf> (accessed on 5 August 2021).
27. Botto, I.L.; Canafoglia, M.E.; Gazzoli, D.; González, M.J. Spectroscopic and Microscopic Characterization of Volcanic Ash from Puyehue-(Chile) Eruption: Preliminary Approach for the Application in the Arsenic Removal. *J. Spectrosc.* **2013**, *2013*, 254517. [\[CrossRef\]](#)
28. Johnson, B.; Turnbull, K.; Brown, P.; Burgess, R.; Dorsey, J.; Baran, A.J.; Webster, H.; Haywood, J.; Cotton, R.; Ulanowski, Z.; et al. In situ observations of volcanic ash clouds from the FAAM aircraft during the eruption of Eyjafjallajökull in 2010. *J. Geophys. Res. Atmos.* **2012**, *117*. [\[CrossRef\]](#)
29. Ball, J.G.C.; Reed, B.E.; Grainger, R.G.; Peters, D.M.; Mather, T.A.; Pyle, D.M. Measurements of the complex refractive index of volcanic ash at 450, 546.7, and 650 nm. *J. Geophys. Res. Atmos.* **2015**, *120*, 7747–7757. [\[CrossRef\]](#)
30. Weber, K.; Eliasson, J.; Vogel, A.; Fischer, C.; Pohl, T.; van Haren, G.; Meier, M.; Grobéty, B.; Dahmann, D. Airborne in situ investigations of the Eyjafjallajökull volcanic ash plume on Iceland and over north-western Germany with light aircrafts and optical particle counters. *Atmos. Environ.* **2012**, *48*, 9–21. [10.1016/j.atmosenv.2011.10.030](https://doi.org/10.1016/j.atmosenv.2011.10.030). [\[CrossRef\]](#)
31. Gudmundsson, M.T.; Thordarson, T.; Höskuldsson, A.; Larsen, G.; Björnsson, H.; Prata, F.J.; Oddsson, B.; Magnússon, E.; Högnadóttir, T.; Petersen, G.N.; et al. Ash generation and distribution from the April–May 2010 eruption of Eyjafjallajökull, Iceland. *Sci. Rep.* **2012**, *2*, 572. [\[CrossRef\]](#)
32. Rose, W.; Gu, Y.; Watson, I.; Yu, T.; Blut, G.; Prata, A.; Krueger, A.; Krotkov, N.; Carn, S.; Fromm, M.; et al. The February–March 2000 Eruption of Hekla, Iceland from a Satellite Perspective. In *Volcanism and the Earth's Atmosphere*; American Geophysical Union: Washington, DC, USA, 2004; pp. 107–132. [\[CrossRef\]](#)
33. Rose, W.I.; Bluth, G.J.S.; Watson, I.M. Ice in Volcanic Clouds: When and Where? In Proceedings of the 2nd International Conference on Volcanic Ash and Aviation Safety, Alexandria, VA, USA, 21–24 June 2004; Office of the Federal Coordinator for Meteorological Services and Supporting Research: Silver Spring, MD, USA, 2004; pp. 3.27–3.33.
34. Durant, A.J.; Shaw, R.A.; Rose, W.I.; Mi, Y.; Ernst, G.G.J. Ice nucleation and overseeding of ice in volcanic clouds. *J. Geophys. Res. Atmos.* **2008**, *113*. [\[CrossRef\]](#)
35. Plu, M.; Scherllin-Pirscher, B.; Arnold Arias, D.; Baro, R.; Bigeard, G.; Bugliaro, L.; Carvalho, A.; El Amraoui, L.; Eschbacher, K.; Hirtl, M.; et al. A tailored multi-model ensemble for air traffic management: Demonstration and evaluation for the Eyjafjallajökull eruption in May 2010. *Nat. Hazards Earth Syst. Sci. Discuss.* **2021**, *2021*, 1–32. [\[CrossRef\]](#)
36. Kox, S.; Bugliaro, L.; Ostler, A. Retrieval of cirrus cloud optical thickness and top altitude from geostationary remote sensing. *Atmos. Meas. Tech.* **2014**, *7*, 3233–3246. [\[CrossRef\]](#)
37. Prata, A.J. Infrared radiative transfer calculations for volcanic ash clouds. *Geophys. Res. Lett.* **1989**, *16*, 1293–1296. [\[CrossRef\]](#)
38. Wen, S.; Rose, W.I. Retrieval of sizes and total masses of particles in volcanic clouds using AVHRR bands 4 and 5. *J. Geophys. Res. Atmos.* **1994**, *99*, 5421–5431. [\[CrossRef\]](#)
39. LeCun, Y.; Denker, J.S.; Solla, S.A. Optimal Brain Damage. In *Advances in Neural Information Processing Systems 2*; Touretzky, D.S., Ed.; Morgan-Kaufmann: San Francisco, CA, USA, 1990; pp. 598–605.
40. Piscini, A.; Carboni, E.; Del Frate, F.; Grainger, R.G. Simultaneous retrieval of volcanic sulphur dioxide and plume height from hyperspectral data using artificial neural networks. *Geophys. J. Int.* **2014**, *198*, 697–709. [\[CrossRef\]](#)
41. Ackerman, S.A. Remote sensing aerosols using satellite infrared observations. *J. Geophys. Res. Atmos.* **1997**, *102*, 17069–17079. [\[CrossRef\]](#)
42. Gray, T.M.; Bennartz, R. Automatic volcanic ash detection from MODIS observations using a back-propagation neural network. *Atmos. Meas. Tech.* **2015**, *8*, 5089–5097. [\[CrossRef\]](#)
43. Schmit, T.J.; Gunshor, M.M.; Menzel, W.P.; Gurka, J.J.; Li, J.; Bachmeier, A.S. Introducing the Next-Generation Advanced Baseline Imager on GOES-R. *Bull. Am. Meteorol. Soc.* **2005**, *86*, 1079–1096. [\[CrossRef\]](#)

-
44. Bessho, K.; Date, K.; Hayashi, M.; Ikeda, A.; Imai, T.; Inoue, H.; Kumagai, Y.; Miyakawa, T.; Murata, H.; Ohno, T.; et al. An Introduction to Himawari-8/9 - Japan's New-Generation Geostationary Meteorological Satellites. *J. Meteorol. Soc. Jpn. Ser. II* **2016**, *94*, 151–183. [\[CrossRef\]](#)
 45. Yang, J.; Zhang, Z.; Wei, C.; Lu, F.; Guo, Q. Introducing the New Generation of Chinese Geostationary Weather Satellites, Fengyun-4. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 1637–1658. [\[CrossRef\]](#)
 46. Meeting on the Intercomparison of Satellite-Based Volcanic Ash Retrieval Algorithms, Madison WI, USA, 29 June–2 July 2015, Final Report. World Meteorological Organization, 2015. Available online: https://web.archive.org/web/20171113102551/http://www.wmo.int/pages/prog/sat/documents/SCOPE-NWC-PP2_VAIntercompWSReport2015.pdf (accessed on 5 August 2021).
 47. Stohl, A.; Prata, A.J.; Eckhardt, S.; Clarisse, L.; Durant, A.; Henne, S.; Kristiansen, N.I.; Minikin, A.; Schumann, U.; Seibert, P.; et al. Determination of time- and height-resolved volcanic ash emissions and their use for quantitative ash dispersion modeling: The 2010 Eyjafjallajökull eruption. *Atmos. Chem. Phys.* **2011**, *11*, 4333–4351. [\[CrossRef\]](#)
 48. Dacre, H.F.; Harvey, N.J.; Webley, P.W.; Morton, D. How accurate are volcanic ash simulations of the 2010 Eyjafjallajökull eruption? *J. Geophys. Res. Atmos.* **2016**, *121*, 3534–3547. [\[CrossRef\]](#)
 49. Plu, M.; Bigeard, G.; Sič, B.; Emili, E.; Bugliaro, L.; El Amraoui, L.; Guth, J.; Josse, B.; Mona, L.; Piontek, D. Modelling the volcanic ash plume from Eyjafjallajökull eruption (May 2010) over Europe: Evaluation of the benefit of source term improvements and of the assimilation of aerosol measurements. *Nat. Hazards Earth Syst. Sci. Discuss.* **2021**, *2021*, 1–24. [\[CrossRef\]](#)
 50. Barnes, L.R.; Schultz, D.M.; Grunfest, E.C.; Hayden, M.H.; Benight, C.C. CORRIGENDUM: False Alarm Rate or False Alarm Ratio? *Weather Forecast.* **2009**, *24*, 1452–1454. [\[CrossRef\]](#)