An aerosol classification scheme for global simulations using the K-means machine learning method

Jingmin Li[1], Johannes Hendricks[1], Mattia Righi[1], Christof G. Beer[1]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

Correspondence to: Jingmin Li (Jingmin.Li@dlr.de)

Abstract. The K-means machine learning algorithm is applied to climatological data of seven aerosol properties from a global aerosol simulation using EMAC-MADE3. The aim is to partition the aerosol properties across the global atmosphere in specific aerosol regimes, mainly for evaluation purposes. K-means is an unsupervised machine learning method with the advantage that an a priori definition of the aerosol classes is not required. Using K-means, we are able to quantitatively define global aerosol regimes, so-called aerosol clusters, and explain their internal properties as well as their location and extension. This analysis shows that aerosol regimes in the lower troposphere are strongly influenced by emissions. Key drivers of the clusters' internal properties and spatial distribution are, for instance, pollutants from biomass burning/biogenic sources, mineral dust, anthropogenic pollution, as well as their mixing. Several continental clusters propagate into oceanic regions as a result of long-range transport of air masses. The identified oceanic regimes show a higher degree of pollution in the northern hemisphere than over the southern oceans. With increasing altitude, the aerosol regimes propagate from emission-induced clusters in the lower troposphere to roughly zonally distributed regimes in the middle troposphere and in the tropopause region. Notably, three polluted clusters identified over Africa, India and eastern China, cover the whole atmospheric column from the lower troposphere to the tropopause region. The results of this analysis need to be interpreted taking the limitations and strengths of global aerosol models into consideration. On the one hand, global aerosol simulations cannot estimate small-scale and localized processes due to the coarse resolution. On the other hand, they capture the spatial pattern of aerosol properties on the global scale, implying that the clustering results could provide useful insights for aerosol research. To estimate the uncertainties inherent in the applied clustering method, two sensitivity tests have been conducted i) to investigate how various data scaling procedures could affect the K-means classification and ii) to compare K-means with another unsupervised classification algorithm (HAC, i.e. Hierarchical Agglomerative Clustering). The results show that the standardization based on sample mean and standard deviation is the most appropriate standardization method for this study, as it keeps the underling distribution of the raw dataset and retains the information of outliers. The two clustering algorithms provide similar classification results, supporting the robustness of our conclusions. The classification procedures presented in this study have a markedly wide application potential for future model-based aerosol studies. ~~A markedly wide application potential of the classification procedure is identified and further aerosol studies are proposed which could benefit from this classification.~~

1 Introduction

Aerosols play an important role in the climate system (Boucher et al., 2013). They influence climate directly by scattering and absorption of solar and terrestrial radiation, as well as indirectly by modifications of cloud properties. The major components of atmospheric aerosols are mineral dust, black carbon (BC) and organic carbon, sulphate, nitrate, ammonium and sea salt. Due to their relatively short residence times, the contributions of these components, their state of mixing as well as the particle size distribution show a large spatial and temporal variability on the global scale (e.g., Lauer and Hendricks, 2006; ~~Li et al. 2009;~~ Mann et al., 2010, 2014; Pringle et al., 2010; Aquila et al., 2011, Sessions et al., 2015, Kaiser et al., 2019). Additionally, their effects on clouds and radiation are highly variable due to the strong dependencies on the physical and chemical properties of the aerosols. This in combination with uncertainties in the current knowledge of key aerosol-related processes makes the quantification of aerosol-climate effects a challenge and results in comparatively large uncertainties in the existing quantifications of the climate impact of anthropogenic aerosols (e.g., Boucher et al. 2013; Myhre et al. 2017, Bellouin et al., 2020).

Global aerosol-climate models equipped with detailed representations of aerosol microphysical and chemical processes are essential tools for the quantification of aerosol-climate effects (e.g., Boucher et al. 1998; Takemura et al. 2005; Stier et al. 2005, 2006; Lauer et al. 2007; Hoose et al. 2008; Righi et al. 2013; Randles et al. 2013; Kipling et al. 2016; Myhre et al. 2017; Bellouin et al., 2020; Righi et al. 2020). During the last decades, considerable attempts have been made by the global aerosol modelling community to develop improved descriptions of aerosol-climate interactions (e.g., ~~Whitby et al. 1997;~~ Ghan and Schwartz, 2007; Boucher et al., 2013; Riemer et al., 2019). Early modelling approaches considered only the mass of aerosol species. However, observations imply that the number, size distribution, and mixing state of aerosols are also critical factors for an accurate representation of aerosol-climate interactions (Albrecht et al. 1989). First attempts of representing the aerosol size distribution and mixing state in global models started at the end of the 20th century (e.g., ~~Whitby et al. 1997;~~ Jacobson 2001). Due to limited computing capacities and the huge computational expenses of global aerosol-climate models, cost effective algorithms have been applied, for instance, the lognormal representations of the aerosol size distribution (e.g., Stier et al. 2005; Lauer et al. 2005; Aquila et al. 2011; von Salzen 2006; Pringle et al., 2010; Kaiser et al. 2019). Recent approaches allow for tracking soluble and insoluble aerosol particle components as well as their mixtures and facilitate the simulation of particle number, mass concentration and size distribution. Beyond the direct radiative impact of aerosols, aerosol-cloud interactions are key processes driving the aerosol climate effects. Hence, parameterizations of aerosol activation in liquid clouds have been established (see Gahn et al., 2011, for a review). In addition, aerosol-induced formation of ice crystals attracts increasing attention (Kanji et al., 2017; Heymsfield et al. 2017). To represent the manifold ice formation pathways induced by a large number of different aerosol types in global aerosol-climate models, the applied microphysical cloud schemes as well as the underlying aerosol sub-models have been further extended (e.g., Lohmann and Kärcher, 2002; Kärcher et al., 2006; Lohmann et al., 2007; Lohmann and Hoose, 2009; Hendricks et al., 2011; Kuebbeler et al., 2014; Righi et al., 2020).

The above examples demonstrate the growing complexity of global aerosol models which, consequently, results in a large number of parameters which describe the aerosol number concentration, size distribution and composition in global models and makes the analysis, evaluation and interpretation of the model results a challenge. This is further complicated by the large spatial and temporal variability of the aerosol properties. Under these circumstances, analysing all relevant variables from a typical global model simulation can become unfeasible. New analysis methods are therefore required to gather information from the huge set of variables and their temporal and spatial variability. A powerful tool to facilitate the analysis of global aerosol model results is the partitioning of the model-simulated aerosol into different groups/clusters, each characterized by specific properties. In the following, these groups will be called aerosol regimes. Information on how these aerosol regimes are distributed in space could be very helpful to obtain a concise but comprehensive view on the complex system of modelled aerosol parameters. Detailed knowledge of the spatial distribution of individual aerosol regimes could be the basis for further analyses and model improvement. For instance, observations within a specific aerosol regime can be combined for evaluating simulation results with regard to this specific aerosol type. Furthermore, model evaluation results based on observations limited in space and time (e.g. aircraft-based field campaigns), could be generalized to a whole aerosol regime covering much larger areas and time periods, assuming that the systematic model biases to be corrected occur nearly homogenously throughout the whole cluster. In addition, knowledge of the properties and spatial extension of aerosol regimes could serve as supportive information for satellite retrieval and for the planning of further field campaigns for aerosol observation.

Previous aerosol classifications have been mainly conducted in the context of observational studies using measurements of aerosol microphysical and optical properties. For example, Groß et al. (2013, 2015) applied classification schemes to identify specific aerosol types and their mixtures based on lidar measurements and satellite data. Their classification procedure follows a tree structure where different aerosol microphysical and optical properties imply different classification branches. This allows to identify complicated vertical stratifications of different aerosol types throughout the atmosphere. Bibi et al. (2016) applied multiple clustering techniques to analyse seasonal differences in prevailing aerosol types at four locations in India. Their classification was based on the analysis of pairs of aerosol optical properties gained from the Aerosol Robotic Network (AERONET) sun photometer measurements. Schmeisser et al. (2017) applied a similar multiple clustering technique to classify aerosol types based on surface-based observations of spectral aerosol optical properties from a global station network. Nicolae et al. (2018) classified six aerosol types using an artificial neural network applied to lidar measurements. The neural network was trained with predefined data from different aerosol types. Applying similar algorithms to global model results using optical aerosol properties to classify aerosol types, however, could be problematic since the optical properties are derived quantities, which are calculated from primary (prognostic) quantities such as aerosol number, size and composition. These calculations also require additional assumptions, usually retrieved from measurements of, e.g. aerosol refractive indices, possibly implying further uncertainties (Dietmüller et al. 2016). Hence new algorithms for aerosol classification based on primary aerosol model parameters would be more appropriate.

In this study, we apply the K-means machine learning clustering algorithm (Hartigan and Wong 1979) for identifying clusters of specific aerosol types in global aerosol simulations. This method partitions *n* samples into *k* clusters in which each sample is assigned to the cluster with the nearest distance to the clusters' centre (or cluster centroid). K-means belongs to the class of unsupervised machine learning algorithms. This is especially useful when the classification criteria are unknown, as in the case of aerosol classification where the specific aerosol characteristics for the predominant regimes are not known a priori. In comparison with supervised classification algorithms which require substantial prior knowledge of classes, an unsupervised classification is relatively easy to use, but it requires the identification and labelling of the resulting clusters after the classification. The common known limitations of K-means include the presence of clusters with equal variances and its sensitivity to outliers. K-means has already been applied in atmospheric research. For instance, it has been successfully used to distinguish clouds and aerosols in CALIOP/CALIPSO observations (Zeng et al. 2019). In this study, we apply the K-means algorithm to global aerosol simulations. The ~~present study aims~~ main goal is to answer the following questions: (1) how can major aerosol regimes be identified in global aerosol simulations? (2) what is the spatial distribution of these regimes? and (3) which aerosol types are dominant in which parts of the world? The K-means method is applied here to identify clusters of different aerosol types in global simulations. The spatial extension of these clusters is quantified. The aerosol properties considered for the clustering process were simulated using the global chemistry-climate model system EMAC (the ECHAM/MESSy Atmospheric Chemistry general circulation model, Jöckel et al. 2010, 2016) equipped with the aerosol microphysical sub module MADE3 (Modal Aerosol Dynamics model for Europe adapted for global applications, third generation, Kaiser et al. 2014, 2019). The aerosol properties analysed here include the mass concentrations of mineral dust, BC, particulate organic matter (POM), sea salt, the sum of aerosol sulphate, nitrate and ammonium (SNA), as well as particle number concentrations in different aerosol size modes. The clustering analysis is conducted separately for the lower troposphere, the mid troposphere and the tropopause region. To quantify potential uncertainties of the clustering procedure, the sensitivity of the results to different methods for scaling the input data is explored. We also provide a comparison of K-means clustering with another unsupervised machine learning clustering algorithm, namely the Hierarchical Agglomerative Clustering (HAC).

The paper is structured as follows: Section 2 describes the model data and the analysis methods in detail. The results of the global clustering procedure are presented in Sect. 3, including separate discussions of the three predefined atmospheric layers. Section 4 provides an uncertainty analysis by testing various sensitivities of the obtained results to methodical aspects, also in view of the limitation and strength of global aerosol models and potential applications of the presented clustering method. ~~Further discussions about the limitation of the applied method and its potential applications are subject of Sect. 4.~~ A summary of the main conclusions as well as an outlook are given in Sect. 5.

2 Data and methods

2.1 Model description and configuration

As a basis for aerosol classification in the present study, we analyse one of the global model simulations of Beer et al. (2020) performed with the global aerosol model EMAC-MADE3. MADE3 simulates nine different aerosol species (sulphate, ammonium, nitrate, the sea salt species sodium and chloride, BC, POM, mineral dust and aerosol water). These nine aerosols species occur in three different internal mixtures (purely soluble particles, mixed particles consisting of an insoluble core with a soluble coating, and particles mainly composed of insoluble material and only very thin soluble coatings) within three size modes (Aitken-, accumulation- and coarse mode). This results in a total of nine aerosol modes. The model considers particle transformations due to coagulation, condensation, gas-particle partitioning and new particle formation. MADE3 was evaluated in detail in past studies and showed a generally good model performance. Kaiser et al (2014) demonstrated the ability of MADE3 to represent the aerosol microphysical processes when compared to a more detailed particle-resolving aerosol model. Kaiser et al. (2019) further demonstrated a good agreement of BC, POM, gaseous species and particle number concentrations simulated with EMAC-MADE3 with various observations. Beer et al. (2020) further extended the model setup of Kaiser et al. (2019) by including an online parameterization for wind-driven dust emissions (Tegen et al., 2002) and performed five model experiments for the time period 2000-2013 in different horizontal and vertical model resolutions. The model results were evaluated by comparison against observational data from the AERONET station network (Holben et al. 1998, 2001) and aircraft-based measurements from the SALTRACE field campaign (Weinzierl et al. 2017). The comparison in Beer et al. (2020) showed that a specific configuration (T63L31Tegen) outperforms the others thanks to its higher resolution and the more detailed representation of dust emission processes. Hence, data from this simulation are used for the clustering analysis in the present study.

For the chosen simulation Beer et al. (2020) applied EMAC in nudged mode, that is, model dynamics were constrained using ECMWF reanalysis data (Dee et al. 2011) including wind divergence and vorticity, temperature, and logarithm of the surface pressure for the years 2000 to 2013. Transient emission data for anthropogenic sources were used to match this simulation period. Anthropogenic emissions were chosen according to the ACCMIP (Atmospheric Chemistry and Climate Model Intercomparison Project; Lamarque et al. 2010) inventory with RCP 8.5 scenario (Riahi et al. 2007, 2011). Biomass burning emissions were taken from the Global Fire Emission Database version 4 (GFED; van der Werf et al. 2017). The wind-driven emissions of mineral dust and sea salt were calculated online for every model time step following the dust parameterization developed by Tegen et al. (2002), and the parameterization of sea spray introduced by Guelle et al. (2001), respectively. As mentioned above, the model was applied at a T63L31 resolution, corresponding to a $1.9^\circ \times 1.9^\circ$ horizontal resolution and 31 vertical hybrid pressure levels covering the vertical range from the surface up to 10 hPa. For a more detailed description of the simulation setup, we refer to Beer et al. (2020). Some important aspects regarding the quality of the aerosol representation in

this simulation, as well as the advantages and disadvantages of global aerosol models in general, are further discussed in Sect. 4.3.

2.2 Data

Seven aerosol parameters extracted from the Beer et al. (2020) simulation are considered for the clustering process: the mass concentrations of mineral dust, BC, POM, sea salt, the sum of the sulphate, nitrate, and ammonium concentration (SNA), as well as Aitken and Accumulation mode particle number concentration N_{akn} and N_{acc} of the combined aerosol species. Using number properties in addition to mass properties is helpful since the number ratio of small to large particles can change even when the total mass stays constant. The number concentrations of coarse mode particles are not taken into account to avoid the duplication of information, since they are strongly correlated with the mass concentration of sea salt and mineral dust, owing to a comparatively small variability in the size distributions of the modelled mineral dust and sea salt particles. Since the size distributions of the modelled Aitken and accumulation modes are more variable, the number concentrations of these particles are considered in addition to the corresponding mass concentrations. The clustering process is intended to identify model grid points with similar climatological mean aerosol parameters, as a basis to classify the global aerosol distribution in different aerosol regimes.

The simulation data from years 2000 to 2013 are first reduced to multi-year (14 years) means to investigate the distribution of climatological aerosol regimes. To account for the vertical variability of aerosol properties, the model data at 31 vertical levels in the terrain following hybrid sigma pressure level are used to calculate values for three atmospheric layers. More specifically, we integrate model level L31-22 for the lower troposphere (up to ~700 hPa), L21-14 for the middle troposphere (~700 to ~300 hPa) and L13-6 for the tropopause region (~300 to ~100 hPa). Note that EMAC vertical levels are ordered top-to-bottom. Due to the terrain following hybrid sigma pressure level concept, these layers only approximately correspond to specific pressure levels. Deviations can occur in particular over elevated terrain (e.g., the Tibetan Plateau) where the pressure is lower in the layer than in other areas. This layer definition in the statistical analysis, however, is more flexible and can easily be adopted to the respective applications.

2.3 Method

The K-means algorithm used in this study is an unsupervised machine learning algorithm which does not require training data based on known and established classifications. It was first introduced by MacQueen (1967) and a more efficient version of K-means was developed by Hartigan and Wong (1979). K-means is a procedure based on the calculation of the squared Euclidean distance (Spencer, 2013). The Euclidean distance describes the distance between two points in the Euclidean space which can be spanned in any integer dimensions. Assuming that p and q are two points in a j-dimensional space, the Euclidean distance $d\,(p, q)$ between p and q is calculated by:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_j - q_j)^2} \quad (1)$$

The K-means method partitions a sample set into a predefined number of clusters (k) using minimization within cluster variances. The basic input of the algorithm is a sample $X = \{x^1, \ldots, x^n\}$ with $x^m = (x_1^m, x_2^m, \ldots x_j^m)$ and $m \in \{1, \ldots, n\}$, where n is the number of data points and j is the number of variable properties. The sample X is grouped into k cluster subsets (S_1, S_2, ... S_k) by minimizing the sum of the variances within each cluster $S_{i=1, \ldots, k}$ as follows:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{X \in S_i} \|x - \mu_i\|^2 \quad (2)$$

where μ_i is the center of cluster S_i (also called cluster centroid) and the term $\|x - \mu_i\|$ is a simplified notation of Eq. (1) describing the Euclidean distances between all samples in x and their cluster center $\mu_{i=1}^{k}$ in j Euclidean dimensions. The *argmin* operator identifies the set of clusters $S_{i=1, \ldots, k}$ which minimizes the total sum of the Euclidean distance. By applying this procedure, each member of X is assigned to a specific cluster. K-means is a stepwise forward iteration process. In the first step, the cluster centroids are assigned randomly and a prototype of the clusters is first estimated using equation (2). Then, in the second step, the cluster centroids are replaced by prototype cluster means. These two steps are iterated until the cluster centroids change only marginally or even stay constant. At this point the corresponding clusters can be regarded as the optimal set of clusters.

~~Choosing the appropriate *k* for K-means algorithm is not straightforward.~~ Selecting the number of clusters k is one of the most challenging tasks in cluster analysis. Researchers developed many different approaches to select k but there is no standard solution which can be generally applied (e.g. Rousseeuw 1987; Sugar and James 2011; Amorim and Hennig, 2015). In this study we use ~~It requires a combination of~~ clustering evaluation metrics in combination with a plausibility check for evaluation of the obtained clusters. ~~expert judgement to evaluate the plausibility of the obtained clusters~~. Two clustering evaluation metrics commonly used are the sum of squared errors (SSE) and the silhouette coefficient (SC; Rousseeuw, 1987). The SSE is the sum of squared errors calculated between all data points and their cluster centre:

$$SSE = \sum_{i=1}^{k} \sum (X - \mu_i)^2 \quad (3)$$

By plotting the SSE as a function of k and looking for the elbow point on the resulting curve, it is possible to identify the level of a mathematical optimization beyond which the further decrease in the error with increasing k is no longer worth the additional computing cost.

The SC is a metric to validate the consistency/similarity within data of clusters and is defined as:

$$SC = \frac{\sum_{i=0}^{n} sc(i)}{n}, \quad (4)$$

with $sc(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$ (5)

where *a(i)* is the averaged distance of sample i to all other samples within a cluster and *b(i)* is the averaged distance of sample i to all samples of its nearest cluster that the sample i is not a part of. SC values range from – 1 to +1, with a higher value

indicating that samples are well matched to the cluster they were assigned to, while they fit poorly to other clusters (Rousseeuw, 1987).

In this study, we apply the K-means clustering algorithm and calculate cluster evaluation metrics using the Python machine learning package scikit-learn (Pedregosa et al. 2011). The individual model grid points of the global simulation (192×96=18432 points at the chosen T63 horizontal resolution) are assigned to k clusters based on the seven simulated aerosol properties as stated in Sect. 2.2. There is no vertical dependency here since the method is applied separately in each of the three atmosphere layers as defined in Sect. 2.2. A common requirement for the K-means algorithm is the standardization of the input dataset, due to the fact that input quantities span different orders of magnitudes and can have different units. Since aerosol mass and number concentrations have different units and are characterized by very different numerical values, each of the individual aerosol properties x_l, $l \in \{1, \dots, j\}$, are standardized to x_l^s ~~xs_l assuming the deviation of the data from their respective mean to follow a Gaussian distribution with zero mean and variance of one~~ by subtracting their respective mean and dividing each value by its respective standard deviation (StandardScaler method in the scikit-learn package):

$$~~xs_l~~ x_l^s = \frac{x_l - \overline{x_l}}{\sigma_l} \quad (6)$$

where x_l^s ~~xsl~~ stands for standardized data, x_l is the original data, $\overline{x_l}$ is the mean and σ_l is the standard deviation of this specific aerosol property l calculated from the whole set of samples. The standardization ensures the comparability of the different aerosol quantities and facilitates evaluating the prominence of individual aerosol properties in the respective regimes. It also avoids clustering due to one dominate species but instead focusing on the connection between the different species.

In summary, we use a standardization method to harmonize the order of magnitude of the different aerosol quantities to ensure comparability and then apply K-means for the aerosol classification tasks. To investigate the robustness of this method, two additional sensitivity tests are conducted in this study. The first test is designed to analyse how data scaling transforms the input aerosol data and how K-means clustering is influenced by different scaling methods. In addition to the standardization method described above, we apply three further data scaling methods for standardizing the aerosol data, namely the MaxMinScaler, the Robustscaler and the Normalizer from the scikit-learn package (Pedregosa et al. 2011) (see Table 1 in Section 4.1). As a further method, we apply the StandardScaler in Eq. (6) to the (base-10) logarithm of the aerosol concentration data to change the data distribution intentionally. A detailed description of this these scaling methods is presented in Sect. 4.1. In the second sensitivity test we compare the results of K-means clustering to those obtained with a different unsupervised machine learning method (HAC), using the StandardScaler standardization. This allows us to investigate whether choosing an alternative clustering algorithm might lead to fundamental differences in the obtained aerosol clusters. Details on this sensitivity test can be found in the Sect. 4.2.

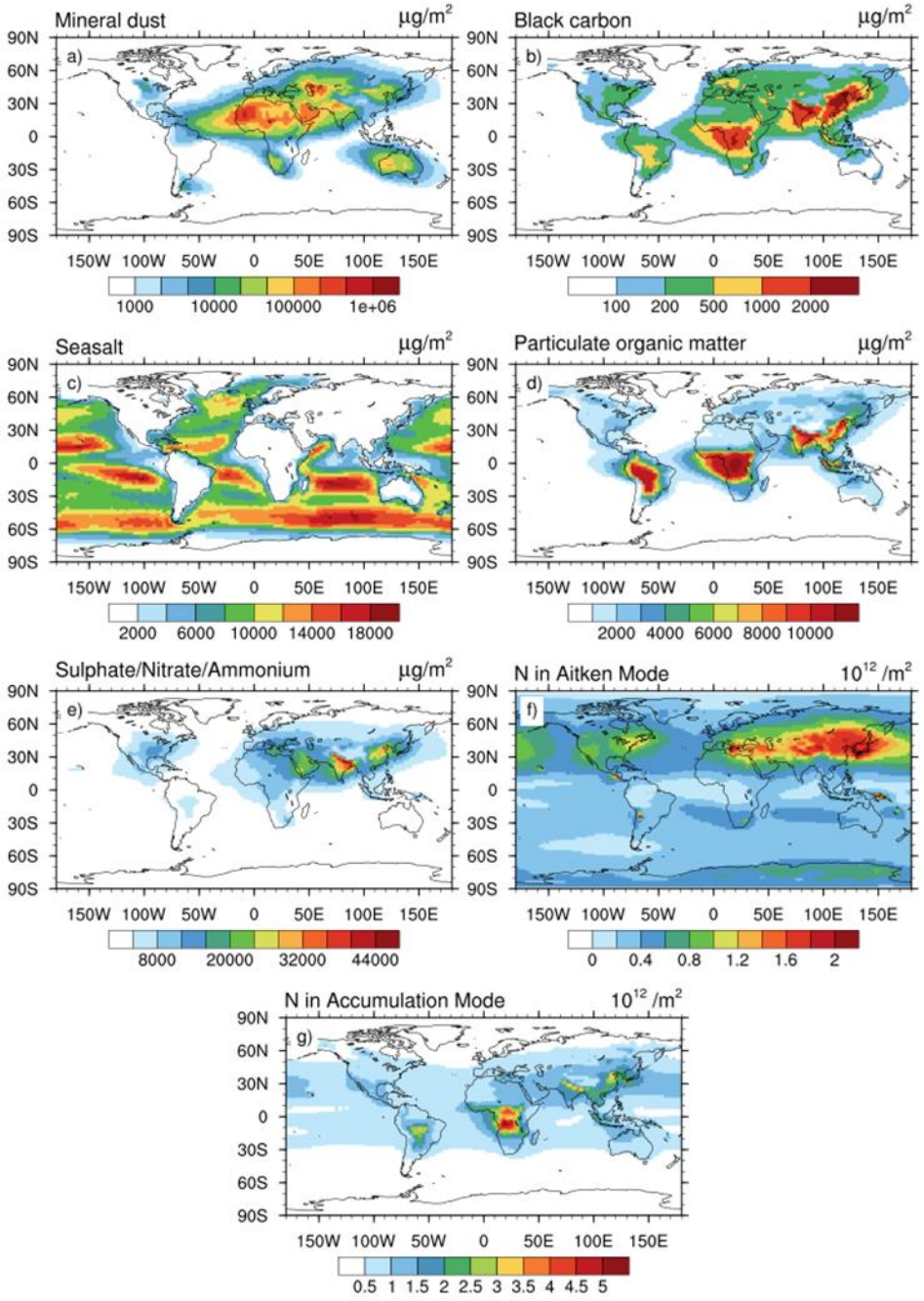


Figure 1: Simulated climatological aerosol properties for the lower troposphere (surface to ~700hPa) including vertically integrated mass concentration of mineral dust (a), BC (b), sea salt (c), POM (d), SNA (e), vertically integrated particle number concentration of the Aitken mode N_{akn} (f) and of the accumulation mode N_{acc} (g).

260

3 Results

In this section we present the results of K-means clustering for global aerosol properties in three atmospheric layers as defined in Sect. 2.2. We focus on 4 aspects: (1) the spatial distribution of the seven individual aerosol properties as inputs for the K-means analyses; (2) the evaluation metrics for the K-means clustering which support the selection of a proper cluster number *k*; (3) the spatial distribution of classified aerosol regimes; and (4) the characteristics identified for each aerosol regime regarding the data distribution of aerosol properties within each class.

The results of the clustering analyses are visualized in this study using global geographical maps of the cluster distributions. In addition, we show so-called box plots which provide additional statistical descriptions of the data distributions for individual aerosol parameters within each cluster. By comparing the data distributions between individual aerosol parameters and regimes we explicitly analyse the characteristics of each regime.

3.1 Lower troposphere clusters

For identifying lower tropospheric clusters, the aerosol mass and number concentrations from the global simulation are vertically integrated from the Earth surface to the model layer which corresponds to about 700 hPa. The resulting spatial distributions are shown in Fig. 1. High mineral dust column masses (up to 1×10^6 μg/m^2) are simulated over the Sahara and in other deserts, while values in other regions are mostly small (Fig.1a). BC column masses are highest in south and east Asia (up to about 3.5×10^3 μg/m^2), due to anthropogenic pollution, and over central Africa (about 2×10^3 μ g/m^2) resulting from intense biomass burning activity (Fig.1b). Peak values of the sea salt column masses over the oceans range between 1×10^4 μg/m^2 and 2×10^4 μg/m^2 (Fig.1c). The pattern of POM columns closely follows that of BC, since the two species share similar emission sources (Fig.1d). Enhanced total masses of sulfate, nitrate, and ammonium (SNA) are noticeable especially over south of the Eurasian continent (up to 5×10^4 μg/m^2) and the Arabian Peninsula (Fig.1e), which could be due to coal burning for energy production (Klimont et al. 2013) especially in the case of India and China. Column integrated numbers of Aitken mode particles, in the following called Aitken mode number columns, are generally high in the Northern Hemisphere, with large values close to strongly polluted areas (Fig.1f), while biomass burning largely contributes to the accumulation mode number column, which is particularly high in prominent biomass burning regions such as Central Africa and South America (Fig.1g). As expected, aerosol mass and number column show a large spatial variation in the lower troposphere, closely following the geographical distribution of the main emission sources. This variability results in a complex pattern of aerosol regimes as shown below.

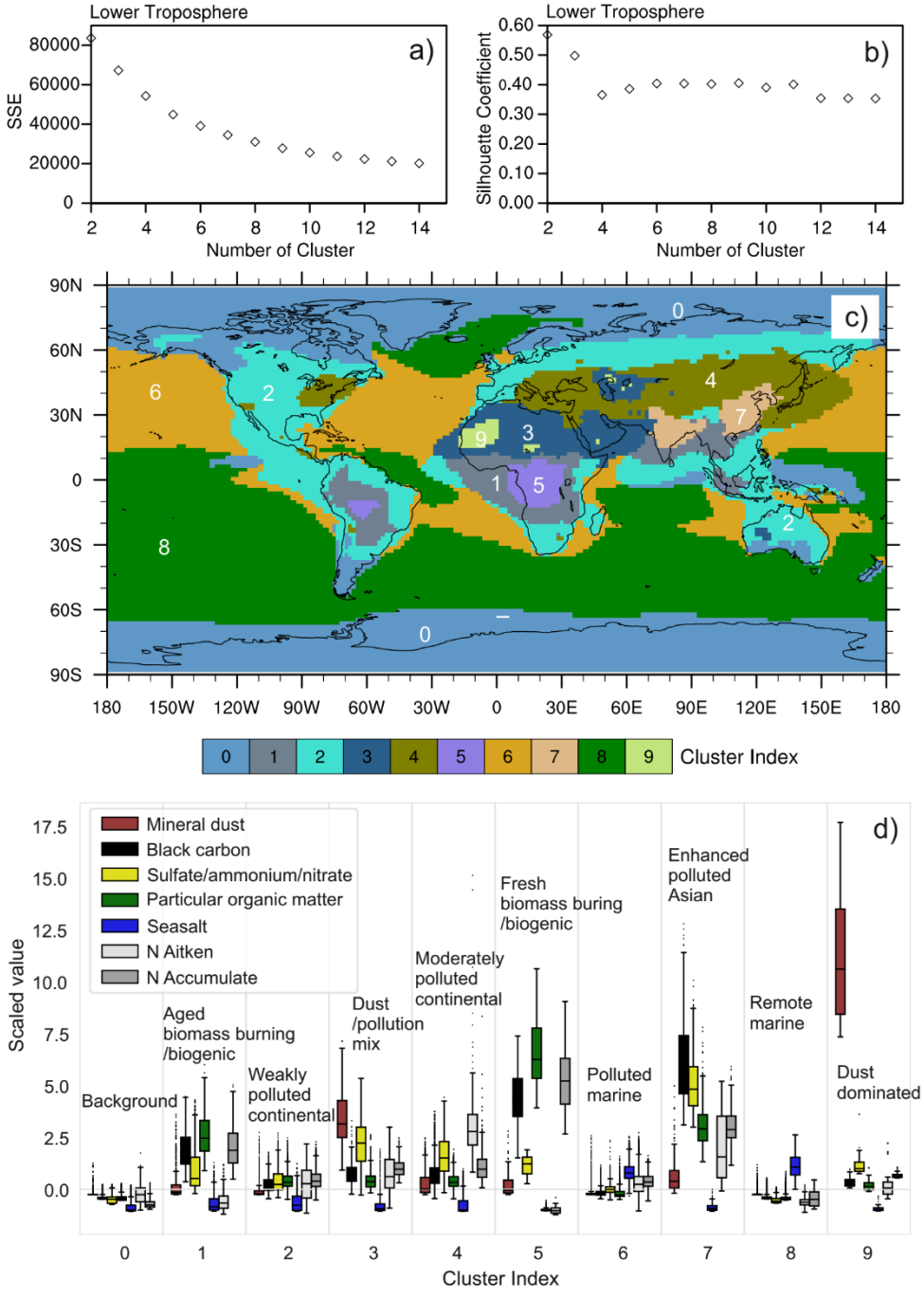


Figure 2: Lower troposphere clustering using K-means. The top panel shows the evaluation metrics SSE (a) and SC~~S~~ (b) vs a *k* range of 2-14. The middle plot (c) highlights the spatial distribution of 10 aerosol regimes for the lower troposphere. The bottom plot (d) shows the data distribution of the 7 considered aerosol properties within the 10 individual aerosol regimes, and cluster names assigned to each cluster based on the analysis of the aerosol data within the respective cluster. The boxplots

describe the distribution of data by displaying 5 statistical quantities that are not outliers: the maximum value (top whisker), 75% quantile, median (top of box), median (middle line in box), 25% quantile (bottom of box) and minimum value (bottom whisker) of standardized aerosol parameters that are not outlies. The black dots are outliers which are defined as the data beyond 2.67σ of a normal distribution.

As explained in Sect. 2.3, K-means classifications are conducted for a range of predefined cluster numbers *k*. The resulting classification is coarse at low *k*, while increasing *k* leads to increased complexity. At some point, however, the added complexity of the K-means classification does not add further information and therefore a further increase of *k* is not useful. Hence, choosing a proper cluster number for the K-means analysis is not straightforward. Here, we use 10 clusters for the lower troposphere based on the K-means evaluation metrics (SSE and SC) and on expert judgement as described above. SSE describes the sum of squared errors from each sample to the respective cluster centre (Eq. 3) and decreases with increasing *k*. For the lower troposphere, SSE decreases rapidly from k=2 up to about k=7 and then more slowly for larger *k* (Fig. 2a). The SC is highest at k=2, decreases between k=2 and k= 4 and reaches a roughly constant level at k=5-11(Fig. 2b). The higher the SC value is, the more similar are the data within the cluster and the more distinct to other clusters. The optimal solution is obtained by minimizing SSE and maximizing the SC. Therefore, taking a balance between small SSE and large SC, we limit the selection of *k* to 9 to 11. The difference between the 9-cluster and the 10-cluster classification is that one oceanic aerosol regime in the 9-cluster classification is further divided into two clusters in the 10-cluster classification. The 11-cluster classification includes a tiny regime which adds little information with respect to the 10-cluster one (Figure S1 in the supplementary material). We therefore choose k=10 for the aerosol classification in the lower troposphere.

The resulting 10 aerosol regimes classified by K-means for the lower troposphere are displayed in Figure 2c. These identified major aerosol classes match well with the expected regimes in this altitude range. Polar aerosols are classified in cluster 0, while oceanic aerosols are roughly divided between Northern and Southern Hemisphere by clusters 6 and 8, respectively. The large forests and savannas of Africa and South America are covered by cluster 5 and cluster 1 including major biogenic and fire aerosol sources (e.g., Dentener et al., 2006). Clusters 9 and 3 cover the main desert regions over Sahara and the Arabian Peninsula. Cluster 9 marks the strong dust emission spots, while cluster 3 represents a kind of “background desert” which shows slight influences by aerosol transported from surrounding areas. The regions characterized by strong anthropogenic pollution (Southern and eastern Asia) are assigned to cluster 7, while regions with moderate and low pollution are covered by cluster 4 and cluster 2, respectively, with the latter often extending to oceanic regions possibly affected by long-range transport of anthropogenic pollution from the continents.

The characterization of the aerosol regimes in the lower troposphere obtained with the K-means method can be further explored and interpreted using the boxplot in Figure 2d. The figure shows the distribution of samples collected within each regime and several statistical metrics, including maximum, 75% quantile, median, 25% quantile and minimum of the standardized aerosol parameters that are not outliers. We recall the use of multi-annual mean sample values and the consideration of column

integrated values in the lower tropospheric column. The dots are outliers that can be ignored for statistical discussion. They are defined by +/- 1.5 times of the interquartile range of the data, which corresponds to data beyond 2.67 sigma of a normal distribution. Note that values on the y-axis are the standardized values (calculated with Eq. 5) but not the absolute value as shown in Fig.1, in order to do a proper classification with K-means and to compare species with different units and scales. All aerosol properties within cluster 0 (polar regions) show lower values than in the other clusters, meaning that this can be considered as aerosol background, as denoted also in Figure 2d. Low values are found also in clusters 6 and 8, with the exception of sea salt, which has enhanced values: we therefore mark these two clusters as oceanic aerosol. Clusters 6 and 8 are very similar, which explains why they are merged into one cluster if a 9-cluster classification is used. The difference between them are the slightly higher values of aerosol properties other than sea salt concentrations within cluster 6, which points to a more polluted marine regime than in cluster 8, which represents remote oceanic regions. Cluster 1 and 5 cover the major forests and savannas in Africa and South America and downwind areas and are characterized by enhanced POM, BC and N_{acc}, which are all typical indicators of strong biomass burning and biogenic activity. The difference between the two clusters is that the enhancement of these quantities is more pronounced in cluster 5 compared to cluster 1. This difference suggests that fresh biomass burning and biogenic aerosol characterize cluster 5, while more aged particles are found in cluster 1 as a result of long-range transport and the subsequent dispersion of the affected air masses in combination with particle wet and dry deposition. Cluster 9 and cluster 3 both show enhanced mineral dust values which agrees with their locations in large deserts or in close proximity to desert regions. Cluster 9 shows much larger mineral dust values and much lower values for the other aerosol properties (in particular SNA and N_{akn}) than cluster 3. This suggests that cluster 9 covers the regions of localized strong dust emissions, while cluster 3 includes dust dominated air masses which are mixed with pollution from other regions. The dominance of BC and SNA in cluster 7 matches well with the large pollution characterizing the south and east Asian regions covered by this cluster. Cluster 7 also shows enhanced POM and number concentrations in both aitken and accumulation modes. We therefore name it the enhanced polluted Asian cluster. Clusters 2 and 4 cover large parts of the Eurasian and American continental regions. Cluster 4 is more polluted than cluster 2, but both are relatively clean compared to other continental clusters nearby (e.g., the strongly polluted Asian regions). We refer to these clusters as moderately polluted continental and weakly polluted continental ~~background~~, respectively. Another important aspect worth noting is that continental aerosol clusters frequently propagate into oceanic regions, showing that this method is also able to capture the long-range transport of pollutants from the emission regions to the relatively clean marine environment. For example, clusters 1, 2, and 3 cover also parts of the middle Atlantic Ocean, cluster 2 also appears over the Pacific Ocean near the west coast of the American continent, and cluster 4 extends over the north western Pacific.

3.2 Middle troposphere clusters

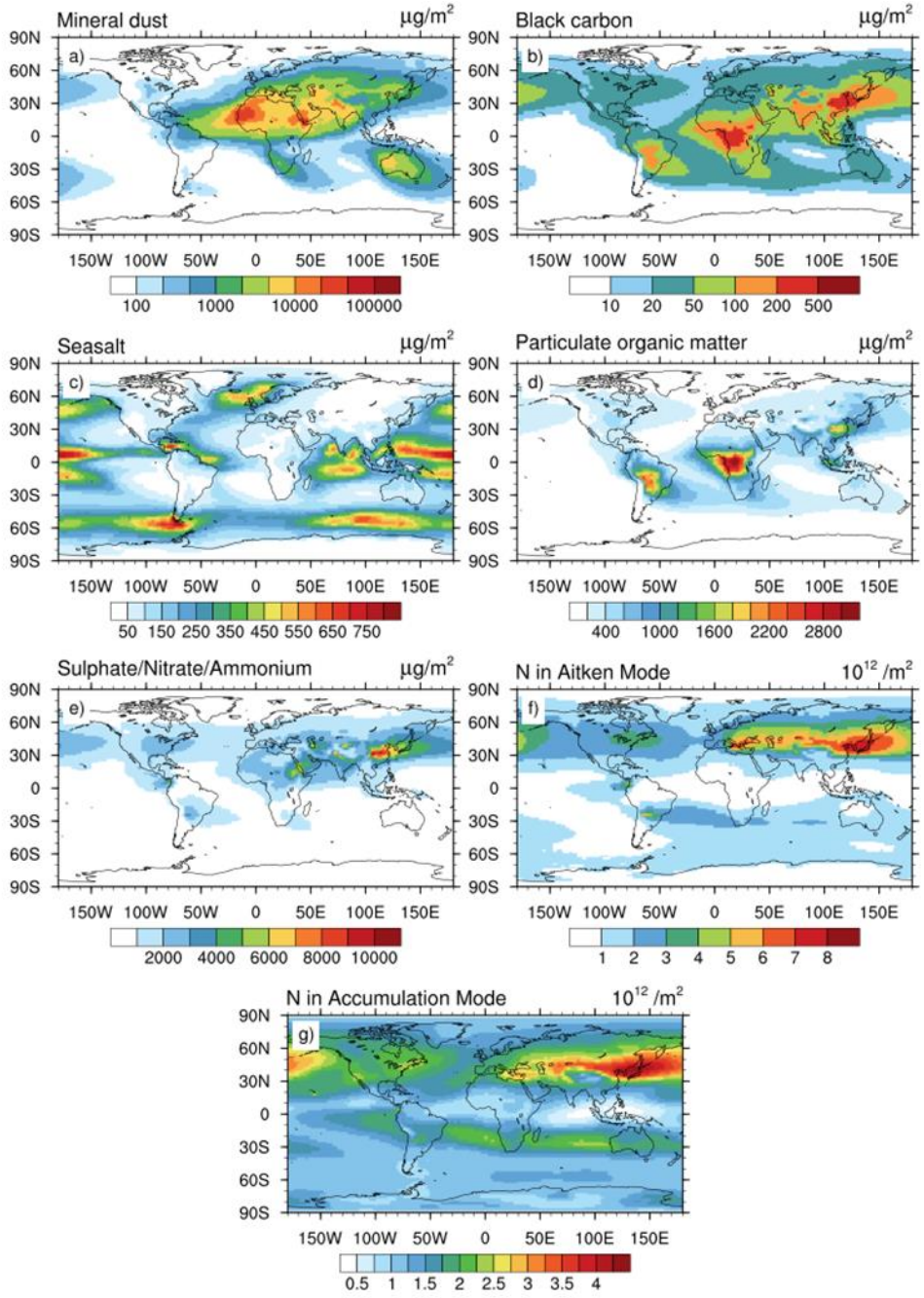


Figure 3: The same as Figure 1 but for the middle troposphere (from ~ 700hPa to ~300hPa).

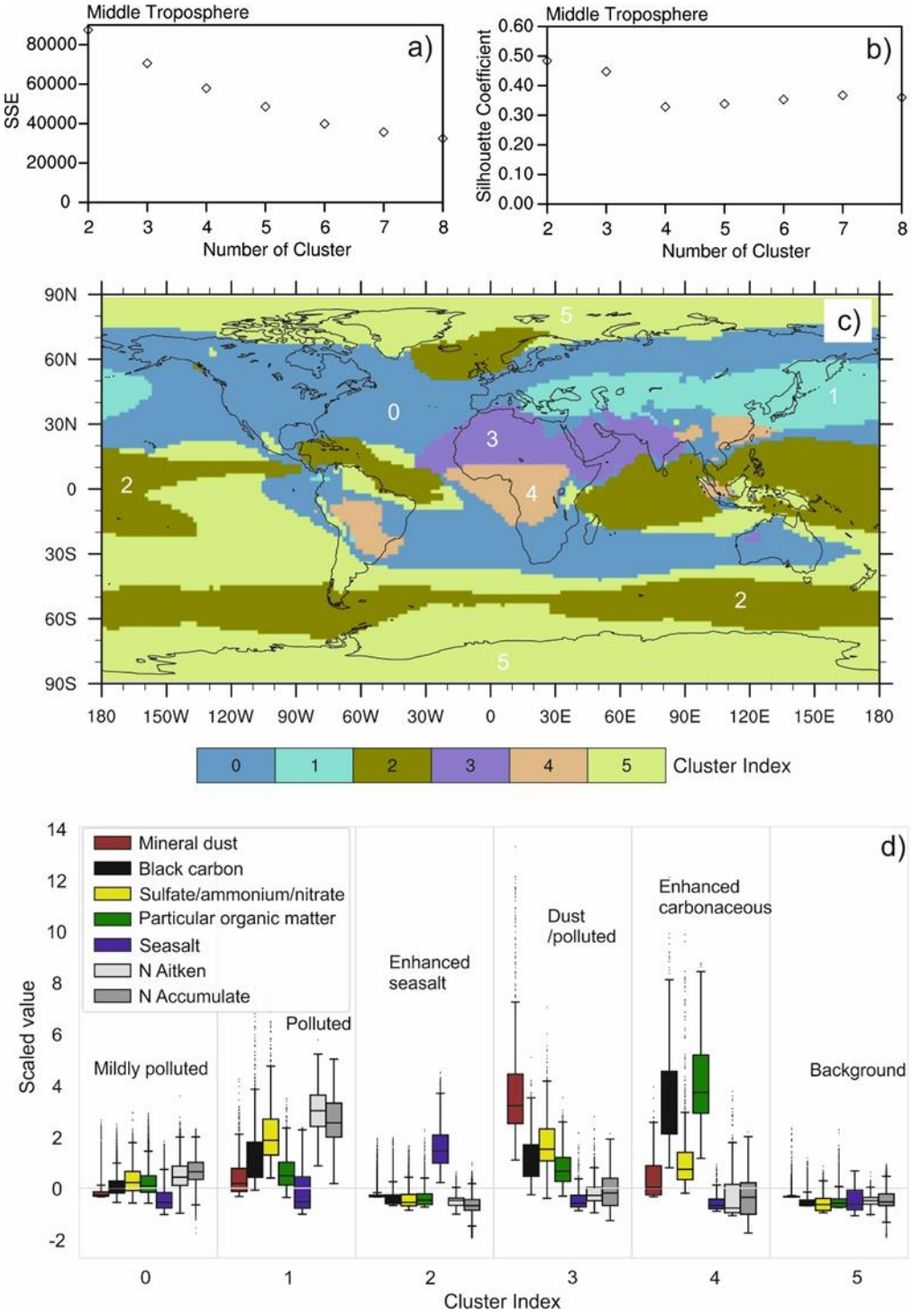
365

The clustering analysis for the middle tropospheric layer uses global aerosol data from about 700 hPa to 300 hPa. As depicted in Fig. 3, this altitude range shows lower values for the column mass and number concentrations (Fig. 1). For example, the column mass of middle troposphere mineral dust (Fig. 3a) ranges from 2 × 103 µg/m2 to 3.4 × 104 µg/m^2 in areas with prominent dust impact, compared to a range of 100 µg/m^2 to 1 × 106 µg/m^2 in the lower troposphere (Fig. 1a). This is caused by the decrease of air density during upward transport, by the dilution of the dust load due to mixing with dust-free air masses as well as by possible sinks due to wet deposition. A similar reduction is also evident in the other aerosol properties. The spatial distribution patterns, however, remain the same between middle troposphere and lower troposphere. However, the overall patterns, in many cases, show a larger spatial extension, caused by long-range transport and dispersion of the respective air masses.

Due to this dispersion, a less complex clustering is required than in the lower troposphere. In general, we can expect *k* to decrease with increasing altitude, due to the more uniform spatial aerosol distributions in the upper atmospheric layers. For the middle troposphere, we evaluated K-means classifications with *k*=2 to *k*=8 using the same metrics as applied above (Fig.4 a and b). As for the lower tropospheric case, SSE decreases with increasing *k*, but more slowly already for *k*≥6. The SC decreases to a minimum for *k*=4 and increases again to a stable level between *k*=6 and *k*=8. The distribution of the major aerosol regimes becomes very robust at *k*=6, while only minor regimes are introduced at higher values which do not show prominent features. We therefore choose a 6-cluster classification for the middle troposphere (See also Figure S2 in the supplementary material).

In the middle troposphere, the aerosol regimes are more zonally uniform than lower down but the lower troposphere has still a very strong influence on the pattern (Fig. 4c). The zonal uniformity particularly occurs in the case of clusters 0, 2 and 5 and appears to be related to the increasing prevalence of zonal wind patterns in the middle troposphere. Clusters 1, 3 and 4, on the other hand, show a stronger influence of the distribution of the emission sources and the transport patterns of the lower troposphere. The statistical analysis of the aerosol properties within each cluster allows to broadly classify the clusters 2 and 5 as middle tropospheric background clusters, and clusters 1, 3, and 4 as middle tropospheric polluted clusters (Fig. 4d). The lowest values of all aerosol properties are found in cluster 5 which can be classified as middle tropospheric background (relatively clean) and covers large fractions of the southern hemispheric oceans and the polar regions. Cluster 2 is characterized by enhanced sea salt values, while the values of other aerosol species remain low as in cluster 5. Hence the cluster includes background air enriched with sea salt due to enhanced wind-driven emissions. Cluster 2 mainly covers the intertropical convergence zone (between 20°S and 20°N) with its strong updrafts and the southern hemispheric storm track area around 60°S, which is also an uplift region between the mid-latitude cell and the polar cell of the main atmospheric circulation pattern. Due to the strong upward transport in these regions, sea salt is lifted from the sea surface to the middle troposphere. Cluster 0 is mainly located in the Northern Hemisphere and above the continents: it is characterized by mildly enhanced BC, SNA, POM, $N_{akn,}$ and N_{acc}. Similar enhancements of some of these aerosol properties are evident in clusters 1, 3, and 4, but with

much larger values. These clusters show similar aerosol characteristics and cover similar regions as their counterparts in the lower troposphere (note however that the algorithm assigns different cluster index numbers for the lower and middle troposphere cases). These three polluted clusters nicely identify three distinct sources: cluster 1 is mostly affected by the strong emission regions in south and east Asia and southern Europe/Mediterranean, cluster 3 presents a mixture of mineral dust and other pollutions sources, with an evident prominence above large deserts, and cluster 4 is an enhanced carbonaceous/biogenic cluster, with significant coverage over the biomass burning and biogenic sources e.g. in South America and Africa. It occurs also over East Asia with its high anthropogenic emissions of carbonaceous particles. Note that the scaled values in Fig. 2d and Fig. 4d should not be compared directly among the different atmospheric layers, because the input data for K-means analyses are scaled individually based on the data within each layer.



410 **Figure 4:** The same as Figure 2 but for the Middle troposphere (from ~ 700hPa to ~300hPa).

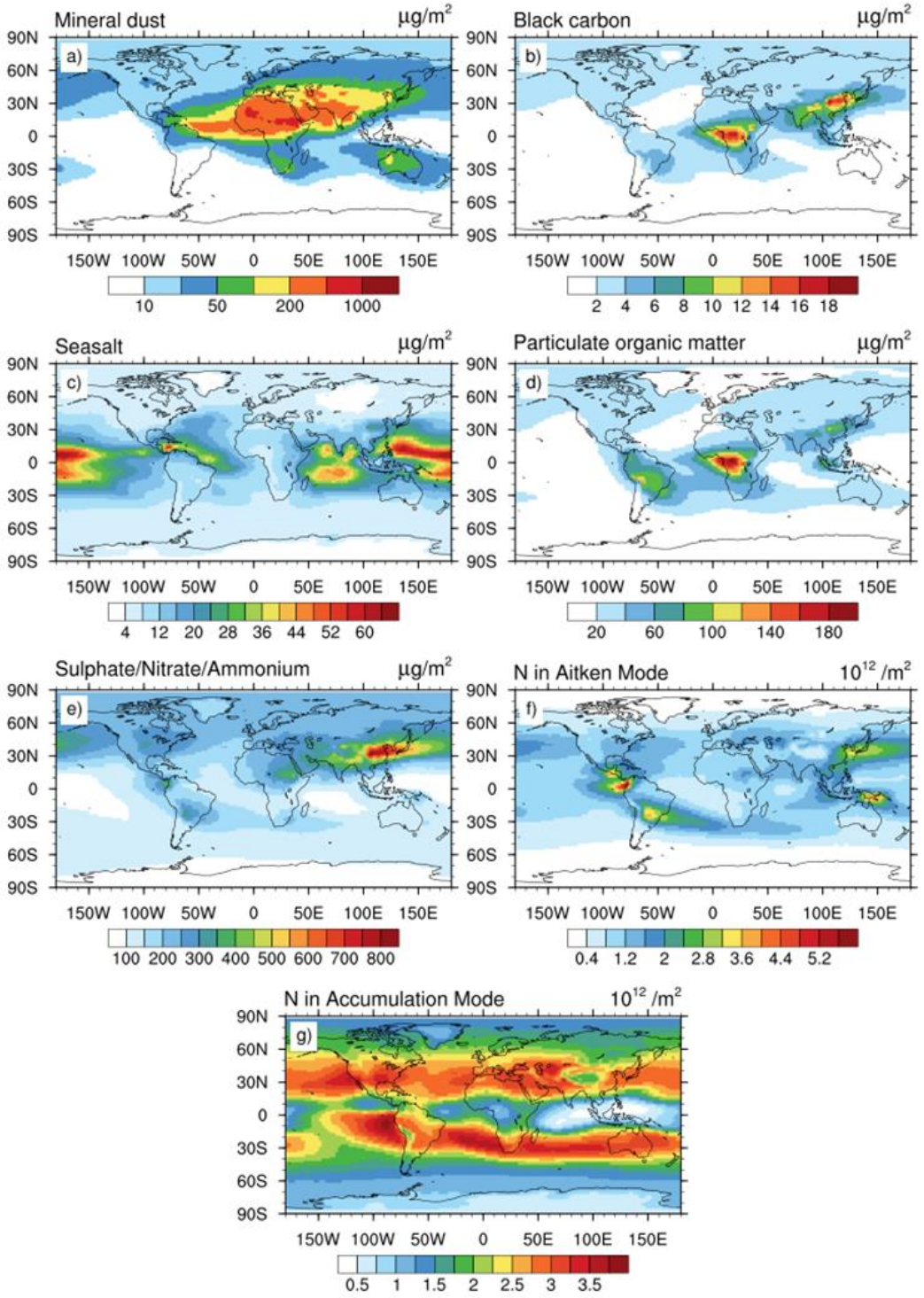


Figure 5: The same as Figure 1 but for the tropopause region (from ~ 300hPa to ~100hPa).

3.3 Tropopause region clusters

The clustering analysis for the tropopause region considers global aerosol data from about 300 hPa to 100 hPa. The degree of spatial dispersion again increases when compared to the lower layers. Therefore, the distributions become more homogeneous than in the middle and lower troposphere (Fig.5). The maximum values of the five aerosol mass columns (mineral dust, BC, sea salt, POM, SNA) are lower in the tropopause region (Fig. 5) than their background value in the lower troposphere (Fig.1). For example, the maximum mineral dust mass column in the tropopause region amounts to about 1×10^3 µg/m^2, which is close to the minimum value of mineral dust in the lower troposphere. Although aerosol mass columns in the tropopause region are generally small and a high degree of dispersion is reached, the spatial patterns for mineral dust, BC, POM and SNA are still related to those in the lower troposphere. This demonstrates that local upward transport of aerosols from the Earth's surface to the tropopause region is efficient in areas showing enhanced dust concentrations. However, this does not fully apply to sea salt, which reaches high values only in the tropics corresponding to regions of strong convection over the oceans into the tropopause region (Fig. 5c). With regard to the aerosol number columns, the effects of vertical and zonal transport appear to be more complex. While the accumulation mode particle number shows a similar behaviour as the mass loadings, the Aitken mode particle number column appears to be strongly influenced by new particle formation in the tropopause region. Hotspots of the particle number occur particularly over regions of enhanced gaseous pollution which provides aerosol precursor gases, such as SO_2, leading to aerosol nucleation and growth favoured by the clean environment of the tropopause region.

As mentioned above and favoured by the homogeneous characteristics of aerosol in the tropopause region shown in Fig. 5, a more simplified clustering can be applied in this layer, reducing *k* to less than 6. Aerosol cluster distributions for a range of different k are shown in Fig. S3 (Supplementary material). The SSE of K-means clustering for the tropopause region (Fig. 6a) shows a similar structure as in the middle troposphere (Fig. 4a), with noticeable convergence from about *k*=6. The SC reaches a maximum for *k*=4 and *k*=5 (Fig. 6b). The combination of these two metrics suggests *k*=5 as the proper choice for the K-means classification for the tropopause region. The resulting 5 clusters are shown in Figure 6c. Large parts of the tropopause region belong to cluster 1, which covers the whole polar regions and most of the southern extra-tropics. The second largest cluster is cluster 2, which covers a large part of the northern extra-tropics and about half of the tropical ocean regions, with the other half mostly covered by cluster 3. Cluster 0 and 4 cover a small portion of the continents including central Africa, the Saharan region as well as tropical and subtropical Asia. Figure 6d highlights the aerosol characteristics for each cluster of the tropopause region. Cluster 1 shows the lowest values for all aerosol properties which suggests to characterize it as tropopause region background. Note that in the polar regions, the pressure levels considered here are mostly located in the stratosphere, and therefore contain comparably clean air. Cluster 3 show similarly low values for all species except for sea salt, which is significantly enhanced due to upward transport in the intertropical convergence zone. Hence, we denote it as the tropopause region enhanced sea salt cluster. The slightly enhanced N_{acc} in cluster 3 relative to the cluster 1 is probably caused by new particle formation. Cluster 2 shows slight increases for all aerosol properties relative to cluster 1, but being still lower than in

the other clusters. We therefore define cluster 2 as the tropopause region mildly polluted cluster. Cluster 0 features strongly increased mineral dust accompanied by slight increases in BC and SNA. Therefore, it can be termed tropopause region dust/polluted cluster. This is also supported by its geographical location over the Sahara and the Middle East where mixtures of desert dust with anthropogenic pollution could be expected. Cluster 4 shows strongly enhanced BC, SNA and POM, and mildly enhanced mineral dust which suggests to term this regime tropopause region polluted/mixed cluster. On the one hand, it is strongly influenced by the biomass burning and biogenic aerosol sources over central Africa and South America. On the other hand, it shows also relevant coverage over East Asia, resulting from the strong pollution sources in these regions. Note that there are many similarities between the aerosol regimes of the tropopause region and the mid troposphere (Fig. 4), especially for clusters 3 and 4, which are largely controlled by efficient updrafts. Hence these clusters correspond also well to lower tropospheric aerosol regimes of similar characteristics occurring in the same regions (Fig. 2).

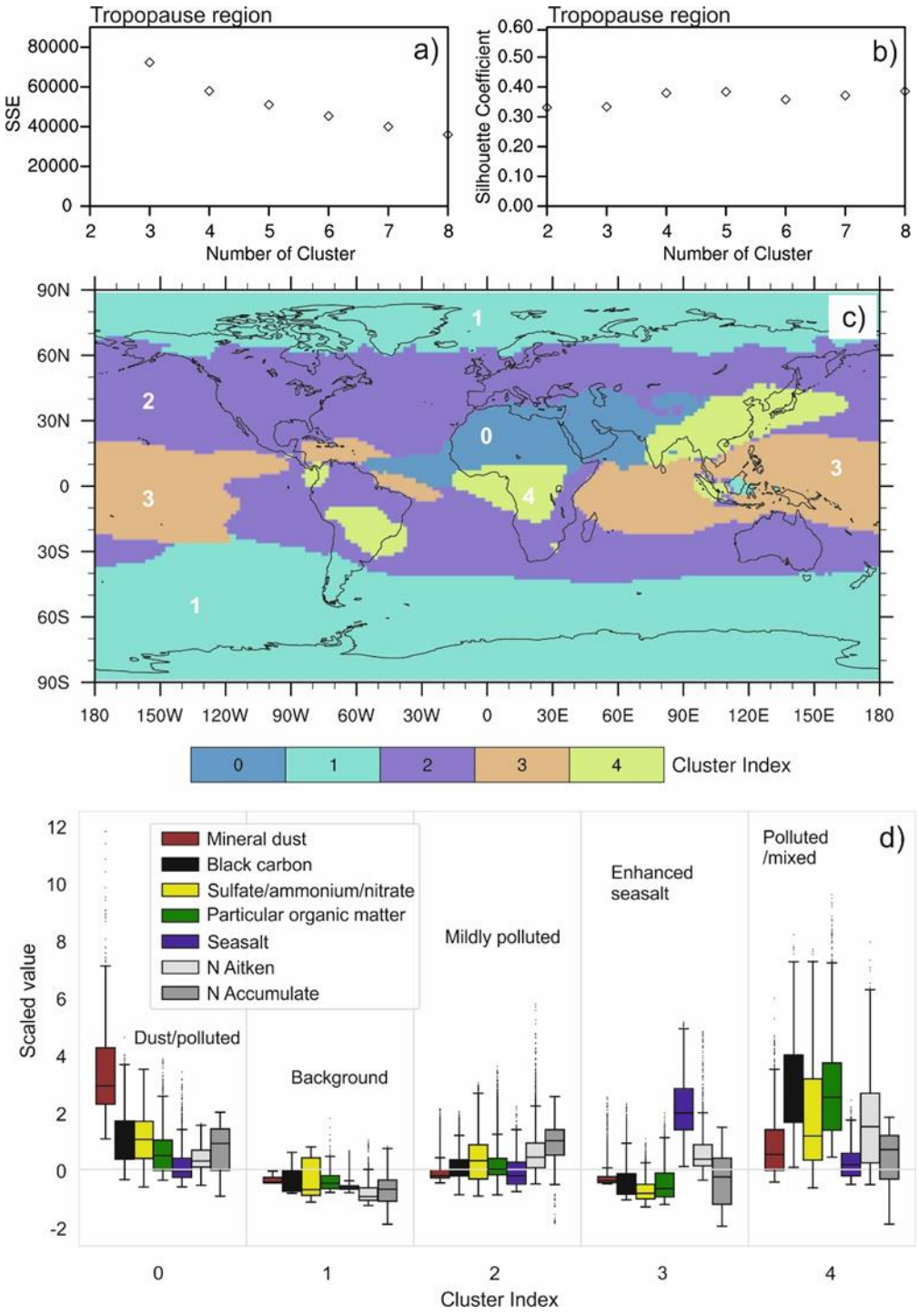


Figure 6: The same as Figure 2 but for the tropopause region (from ~ 300hPa to ~100hPa).

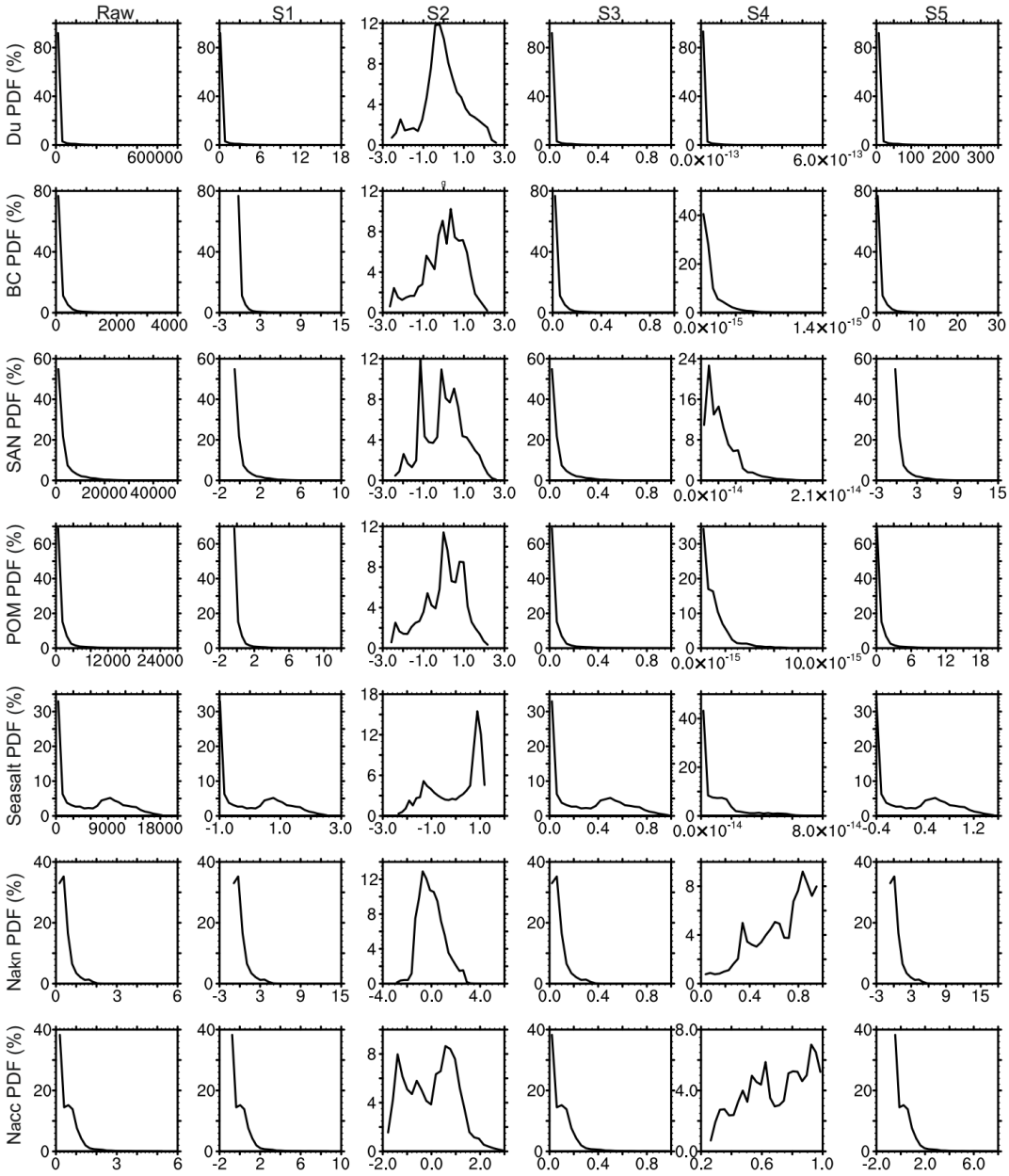
4 Discussion

4.1 Effects of scaling methods on K-means clustering

Table 1: Summary of the different scaling methods applied in this work.

Data scaling	Scikit-learn Function	Definition	Description of the scaled data	Remarks
S1	StandardScaler	Scaling the data of each feature (aerosol property) by subtracting its mean and dividing by its standard deviation.	Scaled data shows a mean value of 0 and a standard deviation of 1.	Reference method chosen in this study.
S2	StandardScaler	Same as S1, but applied to the base-10 logarithm of the input data.	This removes the larger values from the tailed distribution of aerosol properties.	Demonstrates the importance of using original (unchanged) data.
S3	MinMaxScaler	Scaling the data of each feature by subtracting its minimum and dividing by its range.	The values of all scaled properties range between 0 and 1.	Could be used here, but not as suitable as the S1.
S4	Normalizer	Scaling the data on sample (not on feature) by applying Euclidian normalization.	The sum of squared features from a sample (seven aerosol properties) equals to 1.	Not suitable for this study
S5	RobustScaler	Scaling the data of each feature by subtracting its median and dividing by its interquartile range.	The ranges of the scaled properties are larger compared to other methods.	Not suitable for this study

Since the choice of the variance applied for data scaling could potentially have an effect on the clustering, we investigate the influences of different scaling methods on our results in this section. Table 1 summarizes the five tested scaling methods: S1 is the reference standardization method chosen in this study. It is based on Eq. (6). S2 is similar to S1, but applied to the base-10 logarithm of the input data. S3-S5 are alternative methods based on different statistical metrics for standardizing the data. The sensitivity test is applied to the data from the lower troposphere, as this domain is characterized by a larger spatial variability than the mid and upper atmospheric layers, hence more pronounced clustering features can be expected. As an example, we use the 10-cluster distribution. The optimal selection of k could vary among the different standardization approaches, but we choose a fixed value of k to analyze the impact on the results solely due to the standardization method. The selection of an optimal value for k will be addressed again using a different approach in the next section.



475

Figure 7: Probability density functions (PDF) of the seven aerosol properties (rows) derived from their global lower tropospheric distributions in the raw (unscaled) data (first column), and after applying the S1-S5 scaling methods (second to sixth column). The units of the raw (unscaled) values are the same as in Fig.1.

Figure 7 compares the probability density functions (PDF) of the raw input data and the scaled data using the different standardization methods summarized in Table 1. Figures 8 a-e show the distribution of clusters resulting from the differently scaled data and demonstrates how data scaling changes the results of K-means clustering. Based on these results, we can draw the following four conclusions: (1) The standardization which we use for this study (S1) simply scales the values of aerosol properties but it does not change the underling distribution of the raw data (see the first and second column in Fig.7); (2) The most important criterion for K-means data preprocessing is that the data of different properties should be scaled to a comparable range so that they are more or less equally weighted. This is clearly not achieved when using the standardization methods S4 and S5, leading to a large spread in the ranges of scaled data for different aerosol properties (last two columns in Fig. 7). For example, using the S4 method, the maximum scaled value of N_{akn} and N_{acc} is 1.0, while for the other five aerosol properties the maximum values are smaller than 6.0×10^{-13} (Fig.7, fifth column). Similarly, using the S5 method results in much larger values for mineral dust compared to the other aerosol properties (Fig.7, sixth column). As a consequence, the properties with larger values are weighted more strongly in the K-means clustering, leading to a classification largely dominated by these properties (compare Fig. 1a and Fig. 8e). (3) Both the S1 and the S3 methods scale the data to comparable ranges and retains the underlying distribution of the input data, but S1 is more appropriate for this study. For example, sea salt is a natural marine aerosol and its global range of concentration values is relatively narrow, in comparison with the global ranges of other types of aerosols which have both anthropogenic and natural sources or pure natural sources but with locally strong emissions as mineral dust. The maximum values of global sea salt correspond to about 3 standard deviations, while the maximum values of other aerosol properties correspond to about 10-18 standard deviations (Figure 7, second column). This difference is a true feature of the data. Therefore, scaling sea salt and other aerosol properties to the same range of values between zero and one using the S3 method is not suitable for the purpose of this study, since it leads to comparably large weighting of sea salt. The difference in the resulting clusters using the S1 and S3 methods are depicted in Fig. 8: the S3 method (Fig. 8c) results in finer defined clusters over the southern hemispheric ocean regions compared with S1 (Fig. 8a), but at the expense of a less detailed clustering over the continental regions. For the purpose of this study, however, these fine-resolved oceanic clusters are less relevant than a better defined continental clustering. Furthermore, sharply defined southern hemisphere clusters could also be achieved by increasing k using S1 data (Figure S1 in the supplementary material); (4) The 'outliers' in the data distribution are important for aerosol clustering. We tested this by applying the base-10 logarithm to the original (skewed) distribution, resulting in a more gaussian-like distribution (Fig. 7, third column), thus removing the outliers. When applying the K-means algorithm with this method, several polluted clusters vanish (compare Fig.8 a and b). Although the basic structure of clusters is still visible, some important information is not captured with the S2 method. For the purpose of the present work, these high values in the data distribution should not be interpreted as outliers in the general sense, i.e. indicating noise and wrong information, which could hinder K-means clustering, but are rather due to the intrinsically large spatial differences of aerosol properties across the globe and they do provide useful information on the data set. It is also important to recall, that we consider climatological data averaged over a long-term period (14 years), which already excludes unrepresentative high values in the aerosol distribution.

Based on this sensitivity analysis, we conclude that the StandardScaler (S1) standardization method is the most appropriate one for the scope of this study. Although we focus in this section on the lower troposphere, this conclusion holds for the middle troposphere and tropopause region as well (See Figure S4-S7 in the Supplement).

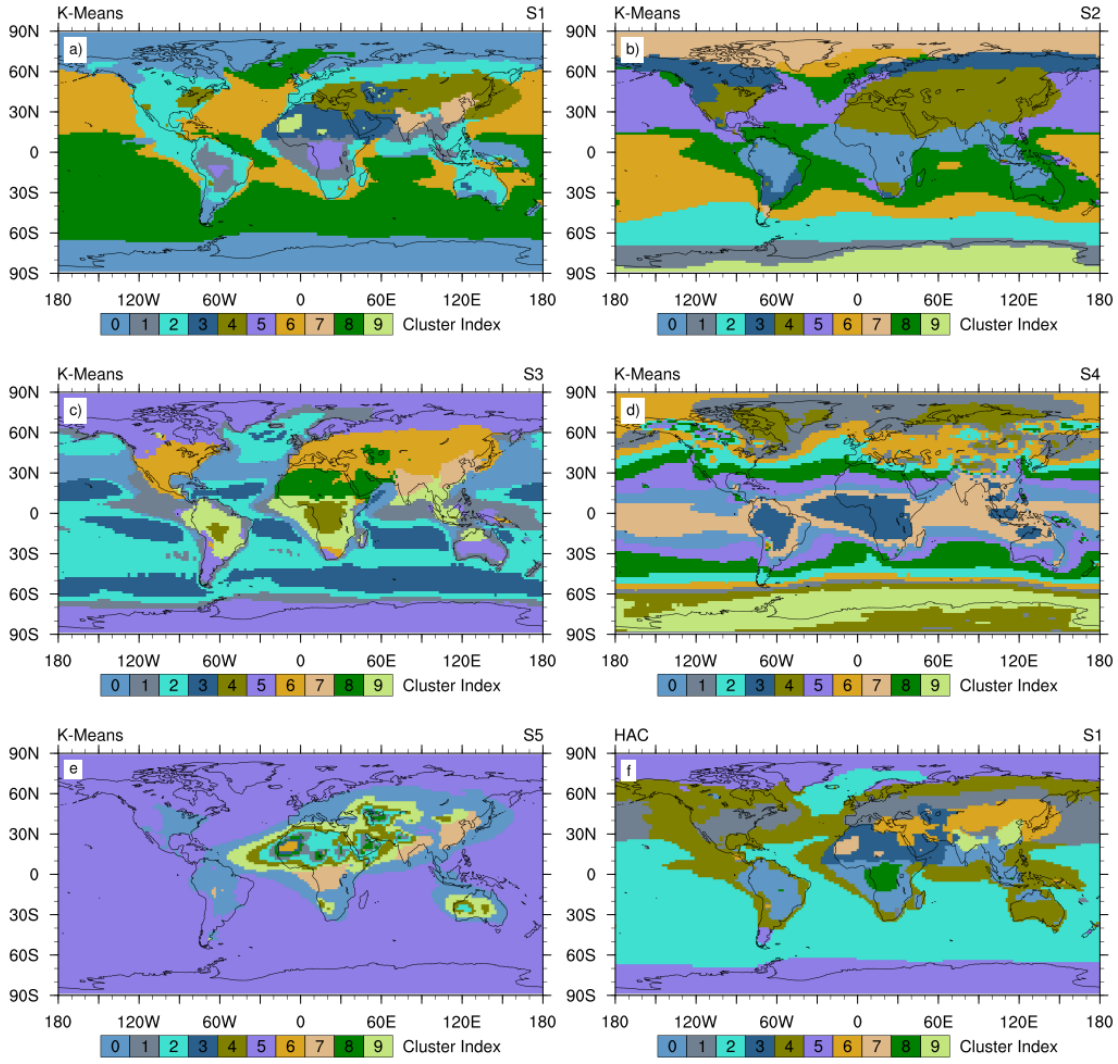


Figure 8: Comparison of K-means 10-cluster distributions based on data scaled with methods S1-S5 (a-e, respectively). Panel (f) shows the HAC clustering method combined with S1 methods.

4.2 Comparison of K-means and HAC clustering

Same as K-means, HAC clustering belongs to the family of unsupervised clustering algorithms. It works with techniques based on hierarchical clustering schemes (e.g. Müllner 2013). More specifically, HAC treats all samples as individual clusters in the first step, then it successively merges pair of clusters which are closest to each other in Euclidean distance, until all samples are grouped into a single cluster. In contrast to K-means, which requires a prescribed number of clusters *k* and separate metrics to evaluate a selection of optimal *k*, HAC shows the hierarchy of clustering along a workflow (the so-called dendrogram), which allows a selection of reasonable cluster number based on this hierarchical structure.

In this section, we compare results of aerosol clustering with HAC and K-means, using the StandardScaler (S1) standardization method and focus on the lower troposphere as an example (additional results for the mid troposphere and tropopause region are provided in the Supplement). The way HAC clustering handles the data points is called linkage. There are different linkage methods such as 'Ward', 'Single', 'Maximum', etc. Here we apply the 'Ward' linkage method for HAC clustering, since it minimizes the sum of squared differences within all clusters and is therefore similar to the K-means apporach. The truncated dendrogram of HAC clustering for the lower tropospheric aerosol is shown in Fig. 9. It demonstrates the path from grouping all samples as individual clusters to one single cluster, and provides insights into the similarities and differences between individual data points or clusters. The distance between two clusters (vertical axis) on the bottom of the hierarchy structure is small but increases as the number of clusters decrease. At a certain level, the dendrogram can be cut in correspondence to the chosen number of clusters. This choice, however, is also subjective and lies in the hand of the investigator. Our selection of 10 clusters is supported by the dendrogram plot which shows a distinct distance between clusters at this level and is also consistent with the selection of 10 clusters for K-means clustering.

The cluster distribution of K-means and HAC shows a good overall agreement but also small differences (Fig.8a, f). We see similarities in the background clusters at the polar regions, the mildly polluted oceanic cluster at northern latitudes and the clean oceanic cluster at southern latitudes, as well as the continental polluted clusters (dust cluster, biogenic cluster, Indian and southeast China cluster). Differences are visible, e.g. in the size of the biogenic cluster over South America, and the size of the mildly polluted continental cluster over the eastern USA. Interestingly, the extent of biogenic clusters over Africa and other continental clusters over Europe and Asia seems to be identical in the two cases. These fine differences in cluster size could be a result of K-means clustering the data by trying to separate samples in groups of equal variances, which HAC does not.

Another aspect to be considered when comparing these two clustering algorithms are the computational expenses. K-means is a fast algorithm. Its computing cost does not scale considerably with sample size or dimensions. HAC has a higher demand on computing time than K-means, especially when the sample size is large. For a sample of size *n*, the computing cost of HAC

scales approximately as n^2 (Dasgupta, 2016; Roy and Chakrabarti, 2017). This is because the hierarchical clustering considers all possible merges at each step, resulting in a rapidly increasing computing time for larger samples. However, HAC features a hierarchy structure (dendrogram) which is more informative and straightforward for deciding on the number of clusters to be used. For this study, both methods provide similar results. Considering further applications of clustering in more complex situations, we chose K-means primarily due to its computational performance.

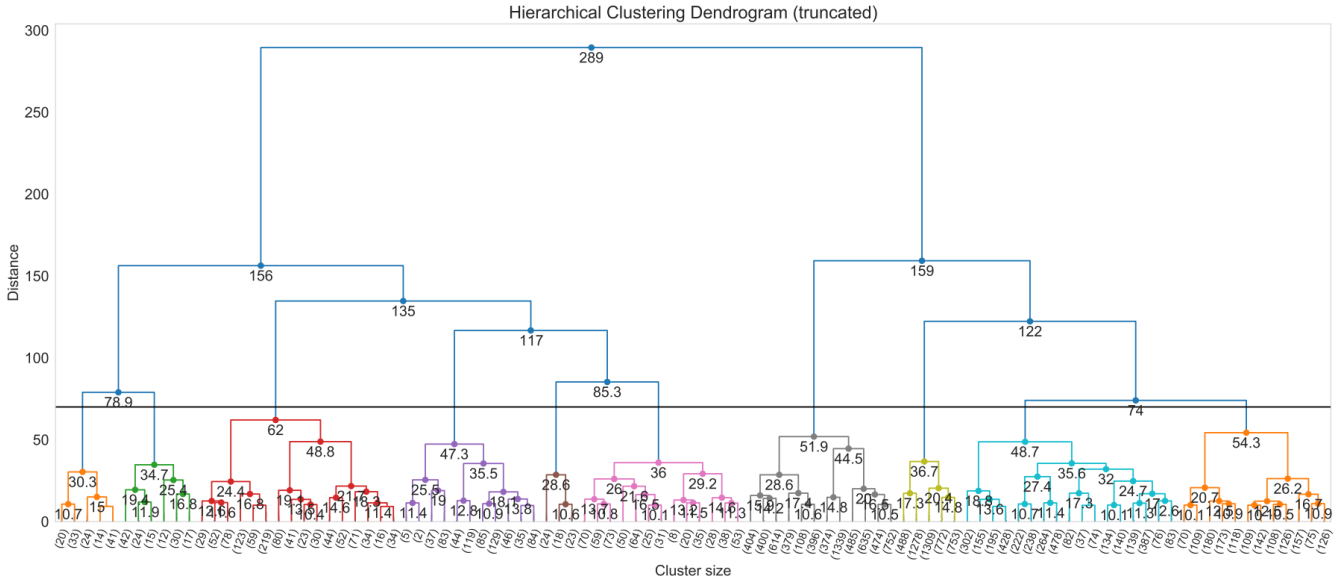


Figure 9: Dendrogram plot of HAC clustering for lower tropospheric aerosols. Since the number of samples (96 latitude × 192 longitude points, resulting in 18432 samples) is too large to be shown on a single plot, the dendrogram is truncated to display only the path of grouping starting from 100 clusters. The values on the horizontal axis represent the number of samples for each branch of these 100 clusters. The horizontal line marks our selection of the cluster number (i.e., k = 10). The distance (y-axis) measure the Euclidean distance between different clusters. The average distance of the merged clusters is highlighted below the clusters.

4.3 Strength and limitation of global aerosol simulation

The major goal of this study is the development of a clustering method to complement classical approaches for analysing and interpreting global aerosols model output. In order to put the demonstration results of the method presented in Sect. 3 in the right context, strengths and limitations of global aerosol simulations are discussed in the following.

Extensive evaluations have been conducted in previous studies to investigate the potential of global aerosol simulations and their limitations (e.g. Textor et al 2006; Lauer et al, 2007; Bauer et al 2008; Koch et al 2009; Mann et al., 2010, 2014; Pringle et al., 2010; Aquila et al 2011; Huneeus et al 2011; Kirkewåg et al. 2013, 2018; He and Zhang, 2014; Koffi et al. 2015; Lee et al 2015; Michou et al 2015; Kaiser et al. 2019). A major deficiency of global aerosol simulations is their inability to resolve small scale and localized processes, largely as a result of the computational challenges and the chemical complexity allowing for only coarse grid resolution in global models. Our clustering analysis is based on data from a global model simulation performed with EMAC-MADE3. The data used has a spatial resolution of about 1.9° × 1.9° in latitude and longitude and can therefore not reproduce smaller-scale features, as for instance aerosol pollution on the scale of specific cities. However, the focus of the present study is the analysis of large-scale global climatological aspects with high relevance for simulating aerosol climate effects. Investigating localized aerosol phenomena and their temporal evolution, which would be of particular relevance for air pollution aspects, is not the intention.

Global aerosol simulations mostly capture the major large-scale spatial patterns of aerosol properties well. For the EMAC-MADE3 model applied here this was demonstrated by Kaiser et al. (2019) and Beer et al (2020). Hence also the clustering results can be expected to show the major large-scale features of the global aerosol distribution. One should keep in mind that for K-means clustering ~~Also considering the fact that~~ the distribution of data is more important than their actual value ~~for K-means clustering~~. Despite the detailed evaluation and improvement of EMAC-MADE3, ~~of this model (Kaiser et al. 2014, 2019) and in particular of the simulation considered here (Beer et al., 2020)~~, some model biases and deficiencies remain and could affect the outcome of the clustering algorithm (Kaiser et al. 2014, 2019; Beer et al. 2020). However, model systematic biases are not necessarily related to wrong data distribution. The model mostly captures the spatial patterns of the aerosol properties, their actual values can be biased. ~~For example, s~~Systematic ~~model~~ biases in model parameterizations and probably also boundary conditions as the considered emission rates (e.g. overestimation/underestimation) cause errors in the absolute values of simulation variables, but these errors ~~are~~ mostly cancelled out when the data are standardized for the K-means analysis. ~~Studies have shown that models generally capture the spatial patterns of aerosol properties quite well but their actual values are biased (Mann et al. 2014; Koffi et al. 2015; Kaiser et al. 2019; Beer et al. 2020).~~ However, simulation biases in the spatial patterns ~~will~~ would change the identified regimes. The extent of such effects ~~this change~~ needs to be further investigated in future studies.

The key advantages of global aerosol simulations are the self-consistent representation of a large number of various aerosol species and properties, the possibility of generating long-term climatological information and future projections, and the global three-dimensional spatial coverage from the surface to the upper atmosphere. This provides a well-suited data base for clustering algorithms. Due to model deficiencies, ~~T~~the clusters derived from the model output could deviate from their appearance in the real atmosphere. However, applying the same algorithm to observational data is not feasible, since no dataset including all relevant chemical and microphysical aerosol properties with global coverage and vertical resolution exists.

Vertically resolved data are available from in-situ aircraft-based measurements, but their geographical coverage is limited and they are often not representative for the ~~on a~~ climatological scale. Satellite data could in principle provide global coverage, but they usually comprise optical aerosol properties, such as aerosol optical depth or aerosol extinction (e.g., Popp et al., 2016). Optical aerosol quantities could be used for classification (e.g. Groß et al, 2015) but the resulting classes do not necessarily reflect the details of aerosol composition and size. In this context, using global model simulation data for classifying global aerosol regimes is an appropriate strategy.

The extensive evaluation performed in the existing global aerosol model studies, considering very large numbers of aerosol-related quantities represented in the simulations, is often difficult to interpret. This, in turn, suggests that new analysis methods, for instance, treating aerosols as groups as presented in this study, are in demand. Although aerosol classification is developed in this study primarily for evaluation purposes, the results of aerosol classification from the global model output potentially provides valuable insights for aerosol research, taking the advantages and limitations of global aerosol simulation into consideration.

4.4 Limitations and potential application of K-means clustering

This study demonstrates the successful application of the K-means algorithm for the classification of global aerosol climatological regimes in model simulation output. It provides quantitative information about the aerosol regimes across the globe and at three altitude ranges, from the surface to the tropopause region. The clustering analysis performed by the algorithm allows to systematically characterize many aerosol properties in a single index, thus facilitating the analysis of the output of global model simulations. This study represents a first attempt to apply the clustering method to global aerosol modelling. However, it has of course limitations and potential for improvements. These are discussed in the following, together with suggestions for possible applications of the presented method.

The K-means method has advantages and disadvantages in performing classification tasks. The advantage is that it does not require prior classification knowledge or training data (Hastie et al., 2009). In cases where no detailed concepts for a pre-definition of aerosol classes based on primary aerosol model parameters can be provided, using K-means is a proper approach. The disadvantage is that the K-means method is sensitive to data variability. Our calculations demonstrated, for instance, that a too high variability resulting from the consideration of temporal variation complicates the K-means clustering. Beyond the analysis of multi-annual means, we attempted to classify global climatological seasonal data which include the variability in the time dimension concerning the four seasons. This attempt resulted in complications in the classification across the four seasons, since the seasonal variations, in many cases, are larger than the differences between the specific clusters, which leads to large changes in the characteristics of the clusters and their spatial extent from season to season. This shows that the K-means method discussed here does not work well for analysing the data variability across time and space simultaneously, as

the interpretation of the resulting classification would be challenging. To overcome this limitation, we removed the variability in the time dimension in this study by considering multi-year averages of the model output, thereby setting a focus on classifying the spatial distribution of long-term climatological aerosol regimes. Possible inter-annual and seasonal variability of aerosol properties could alternatively be discussed on the basis of the climatological regimes analysing the internal temporal changes of aerosol properties within the climatological clusters obtained by K-means.

Despite its limitations the K-means method presented in this study could be ~~is~~ a very helpful tool to analyse and interpret the huge amount of aerosol data generated by global simulations including detailed descriptions of the size-resolved aerosol composition. The method has a wide application potential. Since the algorithm identifies aerosol regimes by minimizing the variance within each cluster, the aerosol properties at different locations within a cluster are similar ~~to each other~~. This implies that aerosols can be treated cluster-wise instead of grid-point-wise, thus reducing the amount of data required to describe the global aerosol population. Possible applications of this method include (but are not limited to) the following:

1. Investigating and correcting model systematic biases using observational data is an important aspect in aerosol model development. However, it is often challenging due to the limited temporal and spatial coverage of observational data. Using the K-means algorithm to identify major aerosol regimes allows to simplify bias-adjustment approaches, since even spatially limited observations within a given cluster can be used to adjust the biases in other regions of that regime. In this context, only systematic model biases which occur nearly homogenously throughout the whole cluster should be addressed, but not purely local model discrepancies. The bias-adjustment for global aerosols remains nevertheless difficult, since it requires a systematic compilation and homogenization of observational aerosol data from different sources, instruments and regions, and requires the consideration of various observational uncertainties. This is planned for a follow-up study.
2. The identified aerosol clusters can be used as first order criteria for satellite retrievals. Some satellite retrieval algorithms (Holzer-Popp et al. 2018; Kahn and Gaitley, 2015) first calculate aerosol optical depth for several pre-defined aerosol types/compositions with top of atmosphere reflectance look-up tables, and then select from the different aerosol types in the atmosphere the best spectral or multi-angular fit between calculated and observed microphysical and optical top of atmosphere reflectance. This is a time-consuming process since a large number of different aerosol types and composition needs to be tested (e.g., 36 or 74 mixtures) without any a-priori pre-selection. By applying the results of the clustering method presented here the characteristics of each aerosol regime could be used to dismiss unrealistic guesses before applying the retrieval algorithm, thus reducing the computing time.
3. Our results could provide data for training other supervised machine learning algorithms. K-means is chosen in this study because a priori definition of aerosol classes is not straightforward since it would require a thorough analysis of the prevailing aerosol regimes in the model output. This however is intended to be achieved with K-means. But if the prevailing aerosol regimes are known from the K-means results, it is possible to prepare training datasets for other

supervised machine learning algorithms for further, more detailed classifications, e.g. using random forest or neural network approaches.

4. The planning of future observational campaigns could benefit from model-based cluster analyses, as they provide useful information on aerosol characteristics in different regimes. Based on this information, campaign planners could easily identify regions of interest regarding specific aerosol properties or types, for example focusing on aerosol from specific sources (e.g. mineral dust from deserts or particles from biomass burning regions).
5. Possible long-term aerosol trends could be analysed by comparing the distribution of clusters calculated for different periods (e.g. pre-industrial, present-day conditions and future scenarios), also providing insights for the validation of climate and air quality measures.

5 Summary and outlook

In this study, we apply the K-means algorithm to classify climatological aerosol regimes across the atmosphere, based on seven primary aerosol properties simulated with the EMAC-MADE3 global aerosol model, primarily for evaluation purposes. These properties include mass concentration of black carbon, mineral dust, sea salt, particulate organic matter, the sulphate/nitrate/ammonium system, and the aerosol number concentrations of the Aitken and accumulation modes. K-means classifies the model data by means of a cluster analysis based on a minimization of the variances, so that data within a respective cluster are similar to each other but different to that in other clusters. K-means ~~has been proven to be a powerful classification tool and~~ is especially useful when prior classification knowledge is not available. We apply K-means to quantitatively identify global aerosol regimes and explain the characteristics of the classified regimes regarding their location, extent, and specific aerosol properties. This study represents the first application of this algorithm for ~~the~~ aerosol classification in global model output ~~of the global aerosol~~. The results show that in the lower troposphere, the aerosol regimes are largely controlled by emissions. Different aerosol clusters are identified, characterized by biomass burning or biogenic activity, mineral dust, anthropogenic pollution, background conditions, as well as a mixture of these different types. Several continental clusters propagate over the oceans due to long range transport of the affected air masses. The algorithm classifies the oceanic regions in two major clusters, with a moderately polluted northern hemisphere and a cleaner southern hemisphere. In the mid troposphere and the tropopause region the aerosol regimes are more zonally uniform than near the surface ~~lower down~~, but the lower troposphere has still a very strong influence on the pattern. Evidences are three polluted clusters occurring over Africa, southern and eastern Asia. Due to efficient vertical dispersion these clusters are present at all altitude levels and show similar characteristics from the surface to the tropopause region.

The above results need to be interpreted keeping the limitation and strength of global aerosol models in mind. Due to the complexity of the processes they simulate in combination with the global, long-term coverage, these models are operable only

with a relatively coarse grid resolution (of the order of 100 Km). Hence the cannot explicitly represent smaller-scale processes, but need to rely on parameterized representations instead. They are, however, a valuable tool to capture the large-scale spatial pattern of aerosol properties, which supports that our results could provide useful insights for aerosol studies.

Two sensitivity tests have been conducted in this study to investigate the robustness of the presented method. Firstly, we investigate on how data scaling influences the K-means classification. By comparing five different data scaling approaches, StandardScaler (S1) standardization is proved to be an appropriate data pre-processing method for this study. Secondly, we explored the differences in classifications purely due to applying an alternative classification algorithm. To this end, the K-means results are compared to the output of another unsupervised classification algorithm (HAC). The results of the classification from both algorithms show good agreement with only small differences in cluster sizes, but the higher computational efficiency of K-means makes it the preferred algorithm for clustering the large data samples resulting from global aerosol model output.

The classification of the global aerosol has a wide spectrum of potential applications. We have suggested several possible future applications that could benefit from this classification scheme. These include identifying model biases and conducting bias-adjustment, preparing training data for other supervised classification algorithms, simplifying satellite retrieval processes, and supporting campaign planning.

Acknowledgements

We are grateful to Dr. Ulrike Brukhardt (DLR, Germany) for her suggestions on an earlier version of this manuscript. We thank Dr. Thomas Popp (DLR, Germany) for his great help on discussing the potential usage of aerosol regimes for satellite retrievals. We are thankful to the developer of Python package Scikit-learn for providing this excellent machine learning package (https://scikit-learn.org/stable/). We are thankful to the 'Jörn's Blog' for sharing scripts for plotting customized Dendrogram (https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/). The model simulations and data analysis for this work used the resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd0080.

Financial Support

This work was supported by the German Federal Ministry for Economic Affairs and Energy - BMWi (project "Digitally optimized Engineering for Services" – DoEfS; contract no. 20X1701B), the DLR space research program (project "Innovative methods for analyzing and evaluating changes in the atmosphere and the climate system" - MABAK), and the DLR transport

program (project “Transport and Climate” – TraK).

Code and data availability

~~Documentation of python package Scikit-learn is available at https://scikit-learn.org/stable/.~~

~~The model simulation data analyzed in this study are available at https://doi.org/10.5281/zenodo.3941462 (Beer, 2020). The cluster analysis code used in this study is available at https://doi.org/10.5281/zenodo.5121180~~

The information on the simulation setup can be found on the zenodo repository for the Beer et al. 2020 paper (https://doi.org/10.5281/zenodo.3941462). The data and scripts used in this study is available at https://zenodo.org/record/5582338.

Author contributions

JL conceived the study, implemented the clustering methods and wrote the paper. JH, MR and CB contributed to conceiving the study, to the interpretation of the results and to the text. CB performed the simulation used in this study.

References

Amorim, R. C. D. and Hennig, C: Recovering the number of clusters in data sets with noise features using feature rescaling factors, Inf. Sci., 324, 126-145, doi: 10.1016/j.ins.2015.06.039, 2015.

Aquila, V., Hendricks, J., Lauer, A., Riemer, N., Vogel, H., Baumgardner, D., Minikin, A., Petzold, A., Schwarz, J. P., Spackman, J. R., Weinzierl, B., Righi, M., and Dall’Amico, M.: MADE-in: a new aerosol microphysics submodel for global simulation of insoluble particles and their mixing state, Geosci. Model Dev., 4, 325-355, doi:10.5194/gmd-4-325-2011, 2011.

Albrecht, B. A.: Aerosols, cloud microphysics, and fractional cloudiness, Science, 245(4923), 1227–1230, doi:10.1126/science.245.4923.1227, 1989.

Bauer, S. E., Wright, D. L., Koch, D., Lewis, E. R., McGraw, R., Chang, L.-S., Schwartz, S. E., and Ruedy, R.: MATRIX (Multiconfiguration Aerosol TRacker of mIXing state): an aerosol microphysical module for global atmospheric models, Atmos. Chem. Phys., 8, 6003–6035, doi:10.5194/acp-8-6003-2008, 2008.

Beer, C. G., Hendricks, J., Righi, M., Heinold, B., Tegen, I., Groß, S., Sauer, D., Walser, A. and Weinzierl, B.: Modelling mineral dust emissions and atmospheric dispersion with MADE3 in EMAC v2.54, Geosci. Model Dev., 13, 4287-4303, doi:10.5194/gmd-13-4287-2020, 2020.

Bellouin, N. and 32 coauthors: Bounding global aerosol radiative forcing of climate change, Rev. Geophys., American Geophysical Union (AGU) 58, e2019RG000660, doi:10.1029/2019RG000660, 2020.

Bibi, H., Alam, K. and Bibi, S.: In-depth discrimination of aerosol types using multiple clustering techniques over four locations in Indo-Gangetic plains, Atmos. Res., 181, 106-114, doi:10.1016/j.atmosres.2016.06.017, 2016.

Boucher, O. and 29 co-authors: Intercomparison of models representing direct shortwave radiative forcing by sulfate aerosols, J. Geophys. Res., 103, 16,979–16,998. doi:10.1029/98JD00997, 1998.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S., Sherwood, S., Stevens, B., and Zhang, X.: Clouds and Aerosols, book section 7, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 571–658, doi:10.1017/CBO9781107415324.016, 2013.

Dee, D. P. and 35 co-authors: The ERA Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

Dentener, F. and 16 co-authors: Emissions of primary aerosol and precursor gases in the years 2000 and 1750 prescribed data-sets for AeroCom, Atmos. Chem. Phys., 6, 4321-4344, doi:10.5194/acp-6-4321-2006, 2006.

Dietmüller., S., Jockel, P., Tost, H., Kunze, M., Gellhorn, C., Brinkop, S., Frömming, C., Ponater, M., Steil, B., Lauer, A. and Hendricks, J.: A new radiation infrastructure for the Modular Earth Submodel System (MESSy, based on version 2.51), Geosci. Model Dev., 9, 2209-2222, doi:10.5194/gmd-9-2209-201, 2016.

Dasgupta, S.: A cost function of similarity-based hierarchical clustering, The 48th Annual ACM SIGACT Symposium, doi: 10.1145/2897518.2897527, 2016.

Ghan, S. J. and Schwartz, E. S.: Aerosol Properties and Processes: A path from field and laboratory measurements to global climate models, Bull. Amer. Meteor. Soc., 88, 1059–1083, doi:10.1175/BAMS-88-7-1059, 2007.

Groß, S., Esselborn, M., Weinzierl, B., Wirth, M., Fix, A., Petzold, A.: Aerosol classification by airborne high spectral resolution lidar observations, Atmos. Chem. and Phys., 13, 2487–2505, doi:10.5194/acp-13-2487-2013, 2013.

Groß, S., Freudenthaler, V., Wirth, M. and Weinzierl, B.: Towards an aerosol classification scheme for future EarthCARE lidar observations and implications for research needs, Atmos. Sci. Let., 16, 77-82, doi:10.1002/asl2.524, 2015.

Guelle, W., Schulz, M., Balkanski, Y., and Dentener, F.: Influence of the source formulation on modeling the atmospheric global distribution of sea salt aerosol, J. Geophys. Res.-Atmos., 106, 27509–27524, doi:10.1029/2001JD900249, 2001

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A k-Means Clustering Algorithm, J. Royal Stat. Soc., 28, 100–108, doi:10.2307/2346830,1979.

Hastie, T., Tibshirani, R., and Friedman, J.: Unsupervised Learning. In: The Elements of Statistical Learning, Springer Series in Statistics, Springer, New York, NY, doi:10.1007/978-0-387-84858-7_14, 2009.

Hendricks, J., Kärcher, B. and Lohmann, U.: Effects of ice nuclei on cirrus clouds in a global climate model, J. Geophys. Res. 116, D18206, doi:10.1029/2010JD015302, 2011.

He, J. and Zhang, Y.: Improvement and further development in CESM/CAM5: gas-phase chemistry and inorganic aerosol treatments, Atmos. Chem. Phys., 14, 9171–9200, doi:10.5194/acp-14-9171-2014, 2014.

Heymsfield, A. J., Krämer, M., Luebke, A., Brown, P., Cziczo, D. J., Franklin, C., Lawson, P., Lohmann, U., McFarquhar, G., Ulanowski, Z., and Tricht, K. V.: Cirrus Clouds, Meteor. Mon., 58, 2.1–2.26, doi.:10.1175/amsmonographs-d-16-0010.1, 2017.

Holben, B. N., Eck, T. F., Slutsker, I., Tanré, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F., Jankowiak, I., and Smirnov, A.: AERONET– A Federated Instrument Network and Data Archive for Aerosol Characterization, Remote Sens. Environ., 66, 1–16, doi:10.1016/S0034-4257(98)00031-5, 1998.

Holzer-Popp, T., Schroedter-Homscheidt, M., Breitkreuz, H., Martynenko, D. and Klüser, L.: Improvements of synergetic aerosol retrieval for ENVISAT, Atmos. Chem. Phys. 8, 7651-7672,2008.

Hoose, C., Lohmann, U., Stier, P., Verheggen, B. and Weingartner, E.: Aerosol processing in mixed-phase clouds in ECHAM5-HAM: Model description and comparison to observations, J. Geophys. Res.-Atmos., 113, D07210, doi:10.1029/2007JD009251, 2008.

Huneeus, N., Schulz, M., Balkanski, Y., Griesfeller, J., Prospero, J., Kinne, S., Bauer, S., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Fillmore, D., Ghan, S., Ginoux, P., Grini, A., Horowitz, L., Koch, D., Krol, M. C., Landing, W., Liu, X., Mahowald, N., Miller, R., Morcrette, J.-J., Myhre, G., Penner, J., Perlwitz, J., Stier, P., Takemura, T., and Zender, C. S.: Global dust model intercomparison in AeroCom phase I, Atmos. Chem. Phys., 11, 7781–7816, doi:10.5194/acp-11-7781-2011, 2011.

Jacobson, M. Z.: GATOR-GCMM: A global- through urban-scale air pollution and weather forecast model 1, Model design and treatment of subgrid soil, vegetation, roads, rooftops, water, sea ice, and snow, J. Geophys. Res., 106, 5385–5401, doi:10.1029/2000JD900560, 2001.

Jöckel, P., Kerkweg, A., Pozzer, A., Sander, R., Tost, H., Riede, H., Baumgaertner, A., Gromov, S., and Kern, B.: Development cycle 2 of the Modular Earth Submodel System (MESSy2), Geosci. Model Dev., 3, 717–752, doi:10.5194/gmd-3-717-2010, 2010.

Jöckel, P., Tost, H., Pozzer, A., Kunze, M., Kirner, O., Brenninkmeijer, C. A. M., Brinkop, S., Cai, D. S., Dyroff, C., Eckstein, J., Frank, F., Garny, H., Gottschaldt, K.-D., Graf, P., Grewe, V., Kerkweg, A., Kern, B., Matthes, S., Mertens, M., Meul, S., Neumaier, M., Nützel, M., Oberländer-Hayn, S., Ruhnke, R., Runde, T., Sander, R., Scharffe, D., and Zahn, A.: Earth System Chemistry integrated Modelling (ESCiMo) with the Modular Earth Submodel System (MESSy) version 2.51, Geosci. Model Dev., 9, 1153–1200, doi:10.5194/gmd-9-1153-2016, 2016.

Kahn, R. A., and Gaitley, B. J.: An analysis of global aerosol type as retrieved by MISR, J. Geophys. Res. Atmos., 120, 4248-4281, doi: 10.1002/2015JD023322, 2015.

Kaiser, J. C., Hendricks, J., Righi, M., Riemer, N., Zaveri, R. A., Metzger, S., and Aquila, V.: The MESSy aerosol submodel MADE3 (v2.0b): description and a box model test, Geosci. Model Dev., 7, 1137–1157, doi:10.5194/gmd-7-1137-2014, 2014.

Kaiser, J. C., Hendricks, J., Righi, M., Jöckel, P., Tost, H., Kandler, K., Weinzierl, B., Sauer, D., Heimerl, K., Schwarz, J. P., Perring, A. E. and Popp, T.: Global aerosol modeling with MADE3(v3.0) in EMAC (basedonv2.53): model description and evaluation, Geosci. Model Dev., 12, 541–579, doi:10.5194/gmd-12-541-2019, 2019.

Kanji, Z. A., Ladino, L. A., Wex, H., Boose, Y., BurkertKohn, M., Cziczo, D. J., and Krämer, M.: Overview of Ice Nucleating Particles, Meteor. Monogr., 58, 1.1–1.33, doi: 10.1175/AMSMONOGRAPHS-D-16-0006.1, 2017.

Kärcher, B., Hendricks, J., and Lohmann, U.: Physically based parameterization of cirrus cloud formation for use in global atmospheric models, J. Geophys. Res.-Atmos., 111, d01205, doi:10.1029/2005JD006219, 2006.

Klimont, Z., Smith, S. J., and Cofala, J.: The last decade of global anthropogenic sulfur dioxide: 2000-2011 emissions, Environ. Res. Lett, 8, 014003, 2013.

Kipling, Z., Stier, P., Johnson, C. E., Mann, G. W., Bellouin, N., Bauer, S. E., Bergman, T., Chin, M., Diehl, T., Ghan, S. J., Iversen, T., Kirkevag, A., Kokkola, H., Liu, X. H., Luo, G., van Noije, T., Pringle, K. J., von Salzen, K., Schulz, M., Seland, O., Skeie, R. B., Takemura, T., Tsigaridis, K. and Zhang, K.: What controls the vertical distribution of aerosol? Relationships between process sensitivity in HadGEM3-UKCA and inter-model variation from AeroCom Phase II, Atmos. Chem. Phys, 16, 2221-2241, doi:10.5194/acp-16-2221-2016, 2016.

Koch, D., Schulz, M., Kinne, S., McNaughton, C., Spackman, J. R., Balkanski, Y., Bauer, S., Berntsen, T., Bond, T. C., Boucher, O., Chin, M., Clarke, A., De Luca, N., Dentener, F., Diehl, T., Dubovik, O., Easter, R., Fahey, D. W., Feichter, J., Fillmore, D., Freitag, S., Ghan, S., Ginoux, P., Gong, S., Horowitz, L., Iversen, T., Kirkevåg, A., Klimont, Z., Kondo, Y., Krol, M., Liu, X., Miller, R., Montanaro, V., Moteki, N., Myhre, G., Penner, J. E., Perlwitz, J., Pitari, G., Reddy, S., Sahu, L., Sakamoto, H., Schuster, G., Schwarz, J. P., Seland, Ø., Stier, P., Takegawa, N., Takemura, T., Textor, C., van Aardenne, J. A., and Zhao, Y.: Evaluation of black carbon estimations in global aerosol models, Atmos. Chem. Phys., 9, 9001–9026, doi:10.5194/acp-9-9001-2009, 2009.

Koffi, B., and 32 coauthors: Evaluation of the aerosol vertical distribution in global aerosol models through comparison against CALIOP measurements: AeroCom phase II results, J. Geophys. Res. Atmos., 121, 7245-7283, doi:10.1002/2015JD0024639, 2015.

Kirkevåg, A., Iversen, T., Seland, Ø., Hoose, C., Kristjánsson, J. E., Struthers, H., Ekman, A. M. L., Ghan, S., Griesfeller, J., Nilsson, E. D., and Schulz, M.: Aerosol–climate interactions in the Norwegian Earth System Model – NorESM1-M, Geosci. Model Dev., 6, 207–244, doi:10.5194/gmd-6-207-2013, 2013.

Kirkevåg, A., Girni, A., Olivié, D., Seland, Ø., Alterskjær, K., Hummel, M., Karset, I. H. H., Lewinschal, A., Liu, X., Makkonen, R., Bethke, I., Griesfeller, J., Schulz, M. and Iversen, T.: A production-tagged aerosol module for Earth system models,OsloAero5.3 – extensions and updates for CAM5.3-Oslo, Geosci. Model Dev., 11, 3945-3982, doi: 10.5194/gmd-11-3945-2018, 2018.

Kuebbeler, M., Lohmann, U., Hendricks, J., and Kärcher, B.: Dust ice nuclei effects on cirrus clouds, Atmos. Chem. Phys., 14, 3027–3046, doi:10.5194/acp-14-3027-2014, 2014.

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., and van Vuuren, D. P.: Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application, Atmos. Chem. Phys., 10, 7017–7039, doi:10.5194/acp-10-7017-2010, 2010.

Lauer, A., and Hendricks, J.: Simulating aerosol microphysics with the ECHAM4/MADE GCM - Part II: Results from a first multiannual simulation of the submicrometer aerosol, Atmos. Chem. Phys., 6, 5495-5513, doi:10.5194/acp-6-5495-2006, 2006.

Lauer, A., Eyring, V., Hendricks, J., Jöckel, P. and Lohmann, U.: Global model simulations of the impact of ocean-going ships on aerosols, clouds, and the radiation budget, Atmos. Chem. Phys., 7, 5061-5079, doi:10.5194/acp-7-5061-2007, 2007.

Lee, Y. H., Adams, P. J., and Shindell, D. T.: Evaluation of the global aerosol microphysical ModelE2-TOMAS model against satellite and ground-based observations, Geosci. Model Dev., 8, 631–667, doi:10.5194/gmd-8-631-2015, 2015.

Lohmann, U. and Kärcher, B.: First interactive simulations of cirrus clouds formed by homogeneous freezing in the ECHAM general circulation model, J. Geophys. Res.-Atmos., 107, D10, doi: 10.1029/2001JD000767, 2002.

Lohmann, U., Stier, P., Hoose, C., Ferrachat, S., Kloster, S., Roeckner, E., and Zhang, J.: Cloud microphysics and aerosol indirect effects in the global climate model ECHAM5-HAM, Atmos. Chem. Phys., 7, 3425–3446, doi:10.5194/acp-7-3425-2007, 2007.

Lohmann, U. and Hoose, C.: Sensitivity studies of different aerosol indirect effects in mixed-phase clouds, Atmos. Chem. Phys., 9, 8917–8934, doi:10.5194/acp-9-8917-2009, 2009.

MacQueen, J. B.: Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07, 1967.

Mann, G. W., Carslaw, K. S., Spracklen, D. V., Ridley, D. A., Manktelow, P. T., Chipperfield, M. P., Pickering, S. J. and Johnson, C. E.: Description and evaluation of GLOMAP-mode: a model global aerosol microphysics model for the UKCA composition-climate model, Geosci. Model Dev., 3, 519-551, doi:10.5194/gmd-3-519-2010, 2010.

Mann, G. W. and 51 coauthors: Intercomparison and evaluation of global aerosol microphysical properties among AeroCom models of a range of complexity, Atmos. Chem. Phys., 14, 4679-4713, doi:10.5194/acp-14-4679-2014, 2014.

Michou, M., Nabat, P., and Saint-Martin, D.: Development and basic evaluation of a prognostic aerosol scheme (v1) in the CNRM Climate Model CNRM-CM6, Geosci. Model Dev., 8, 501–531, doi:10.5194/gmd-8-501-2015, 2015.

Müllner, D.: Fastcluster: Fast hierarchical, agglomerative clustering Routines for R and Python, J. Stat. Softw. 53, 1-18, doi: 10.18637/jss.v053.i09, 2013.

Myhre, G., Aas, W., Cherian, R., Collins, W., Faluvegi, G., Flanner, M., Forster, P., Hodnebrog, O., Klimont, Z., Lund, M., Mulmenstadt, J., Lund Myhre, C., Olivie, D., Prather, M., Quaas, J., Samset, B., Schnell, J., Schulz, M., Shindell, D., Skeie,

R., Takemura, T. and Tsyro, S.: Multi-model simulations of aerosol and ozone radiative forcing due to anthropogenic emission changes during the period 1990-2015, Atmos. Chem. Phys., 17, 2709-2720, doi:10.5194/acp1727092017, 2017.

Nicolae, D., Vasilescu, J., Talianu, C., Binietoglou, I., Nicolae, V., Andrei, S. and Antonescu, B.: A neural network aerosol-typing algorithm based on lidar data, Atmos. Chem. Phys., 18, 14511-14537, doi:10.5194/acp-18-14511-2018, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay É.: Scikit-learn: Machine Learning in Python, JMLR 12, 2825-2830, 2011.

Popp, T. and 31 coauthors: Development, production and evaluation of aerosol climate data records from European Satellite observations (Aerosol_cci), Remote Sens. 8, 421, doi: 10.3390/rs8050421, 2016.

Pringle, K. J., Tost, H., Message, S., Steil, B., Giannadaki, D., Nenes, A., Fountoukis, C., Stier, P., Vignati, E. and Lelieveld, J.: Description and evaluation of GMXe: a new aerosol submodel for global simulations (v1), Geosci. Model Dev. 3, 391-412, doi:10.5194/gmd-3-391-2010, 2010.

Randles, C. A. and 30 co-authors: Intercomparison of shortwave radiative transfer schemes in global aerosol modeling: Results from the AeroCom Radiative Transfer Experiment, Atmos. Chem. Phys., 13, 2347–2379, doi:10.5194/acp-13-2347-2013, 2013.

Riahi, K., Grübler, A., and Nakicenovic, N.: Scenarios of longterm socio-economic and environmental development under climate stabilization, Technol. Forecast. Soc. Change, 74, 887–935, doi: 10.1016/j.techfore.2006.05.026, 2007.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5 – A scenario of comparatively high greenhouse gas emissions, Clim.Change, 109, 33–57, doi:10.1007/s10584-011-0149-y, 2011.

Riemer, N., Ault, A. P., West, M., Craig, R. L. and Curtis, J. H.: Aerosol mixing state: measurements, modeling and impacts, Rev. Geophys., 57, 187-249, doi:10.1029/2018RG000615, 2019.

Righi, M., Hendricks, J. and Sausen, R.: The global impact of the transport sectors on atmospheric aerosol: simulations for year 2000 emissions, Atmos. Chem. Phys., 13, 9939-9970, doi:10.5194/acp-13-9939-2013, 2013.

Righi, M., Hendricks, J., Lohmann, U., Beer, C. J., Hahn, V., Heinold, B., Heller, R., Krämer M., Ponater, M., Rolf, C., Tegen, I. and Voigt, C.: Coupling aerosols to (cirrus) clouds in the global EMAC-MADE3 aerosol-climate model, Geosci. Model Dev., 13, 1635-1661, doi:10.5194/gmd-13-1635-2020, 2020.

Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Comput. Appl. Math., 20, 53-65, doi:10.1016/0377-0427(87)90125-7, 1987.

Roy, S. G. and Chakrabarti, A.: Chapter 11 – A novel graph clustering algorithm based on discrete-time quantum random walk, Quantum Inspired Computational Intelligence, Research and Applications, 361-389, doi: 10.1016/B978-0-12-804409-4.00011-5, 2017.

Schmeisser, L., Andrews, E., Ogren, J. A., Sheridan, P., Jefferson, A. Sharma, S., Kim, J. E., Sherman, J. P., Sorribas, M., Kalapov, I., Arsov, T., Angelov, C., Mayol-Bracero, O. L., Labuschagne, C., Kim, S.-W., Hoffer, A., Lin, N.-H., Chia, H.-

P., Bergin, M., Sun, J., Liu, P. and Wu, H.: Classifying aerosol type using in situ surface spectral aerosol optical properties, Atmos. Chem. Phys, 17, 12097-12120, doi:10.5194/acp-17-12097-2017, 2017.

Sessions, W. R. and 24 coauthors: Development towards a global operational aerosol consensus: basic climatological characteristics of the International Cooperative for Aerosol Prediction Multi-Model Ensemble (ICAP-MME), Atmos. Chem. Phys., 15, 335-362, doi:10.5194/acp-15-335-2015, 2015.

Spencer, N. H.: 5.4.5 Squared Euclidean Distances, Essentials of Multivariate Data Analysis, CRC Press, p. 95, ISBN 9781466584792, 2013.

Stier, P., Feichter, J., Kinne, S., Kloster, S., Vignati, E., Wilson, J., Ganzeveld, L., Tegen, I., Werner, M., Balkanski, Y., Schulz, M., Boucher, O., Minikin, A. and Petzold, A.: The aerosol-climate model ECHAM5-HAM, Atmos. Chem. Phys. 5, 1125-1156, 2005.

Stier, P., Feichter, J., Kloster, S., Vignati, E. and Wilson, J.: Emission-induced nonlinearities in the global aerosol system: Results from the ECHAM5-HAM aerosol-climate model, J. Clim., 19, 3845–3862, doi:10.1175/JCLI3772.1, 2006.

Sugar, G. A. and James, G. M.: Finding the number of clusters in a Dataset, J. Am. Stat. Assoc., 98, 750-763, doi: 10.1198/016214503000000666, 2011.

Takemura, T., Nozawa, T., Emori, S., Nakajima, T. Y. and Nakajima, T.: Simulation of climate response to aerosol direct and indirect effects with aerosol transport-radiation model, J. Geophys. Res. - Atmos., 110, D02202, doi:10.1029/2004JD005029, 2005.

Textor, C., Schulz, M., Guibert, S., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Feichter, H., Fillmore, D., Ghan, S., Ginoux, P., Gong, S., Grini, A., Hendricks, J., Horowitz, L., Huang, P., Isaksen, I., Iversen, I., Kloster, S., Koch, D., Kirkevåg, A., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Liu, X., Montanaro, V., Myhre, G., Penner, J., Pitari, G., Reddy, S., Seland, Ø., Stier, P., Takemura, T., and Tie, X.: Analysis and quantification of the diversities of aerosol life cycles within AeroCom, Atmos. Chem. Phys., 6, 1777–1813, doi:10.5194/acp-6-1777-2006, 2006.

Tegen, I., Harrison, S. P., Kohfeld, K., Prentice, I. C., Coe, M. and Heimann, M.: Impact of vegetation and preferential source areas on global dust aerosol: Results from a model study, J. Geophys. Res. - Atmos., 107, 4576, doi:10.1029/2001JD000963, 2002.

van der Werf, G. R., Randerson, J. T., Giglio, L., van Leeuwen, T. T., Chen, Y., Rogers, B. M., Mu, M., van Marle, M. J. E., Morton, D. C., Collatz, G. J., Yokelson, R. J., and Kasibhatla, P. S.: Global fire emissions estimates during 1997–2016, Earth Syst. Sci. Data, 9, 697–720, doi:10.5194/essd-9-697-2017, 2017.

Von Salzen, K.: Piecewise log-normal approximation of size distributions for aerosol modelling, Atmos. Chem. Phys., 6, 1351–1372, doi:10.5194/acp-6-1351-2006, 2006.

Weinzierl, B., Ansmann, A., Prospero, J. M., Althausen, D., Benker, N., Chouza, F., Dollner, M., Farrell, D., Fomba, W., Freudenthaler, V., Gasteiger, J., Groß, S., Haarig, M., Heinold, B., Kandler, K., Kristensen, T. B., Mayol-Bracero, O. L., Müller, T., Reitebuch, O., Sauer, D., Schäfler, A., Schepanski, K., Spanu, A., Tegen, I., Toledano, C., and Walser, A.: The

Saharan Aerosol Long-Range Transport and Aerosol–Cloud-Interaction Experiment: Overview and Selected Highlights, B. Am. Meteorol. Soc., 98, 1427–1451, doi:10.1175/BAMS-D-15-00142.1, 2017.

Whitby, E. R. and McMurry, P. H.: Modal aerosol dynamics modelling, Aerosol Sci. Tech., 27, 673-688, doi: 10.1080/02786829708965504, 1997.

Zeng, S., Vaughan, M., Liu, Z., Trepte, C., Kar, J., Omar, A., Winker, D., Lucker, P., Hu, Y., Getzewich, B. and Avery, M.: Application of high-dimensional fuzzy k-means cluster analysis to CALIOP/CALIPSO version 4.1 cloud-aerosol discrimination, Atmos. Meas. Tech., 12, 2261-2285, doi:10.5194/amt-12-2261-2019, 2019.