Identifying source regions of air masses sampled at the tropical high-altitude site of Chacaltaya using WRF-FLEXPART and cluster analysis

Diego Aliaga[1], Victoria A. Sinclair[1], Marcos Andrade[2,3], Paulo Artaxo[4], Samara Carbone[5], Evgeny Kadantsev[1,6], Paolo Laj[1,7], Alfred Wiedensohler[8], Radovan Krejci[9], and Federico Bianchi[1]

[1]Institute for Atmospheric and Earth System Research /Physics, Faculty of Science, University of Helsinki, Helsinki, 00014, Finland
[2]Laboratory for Atmospheric Physics, Institute for Physics Research, Universidad Mayor de San Andrés, La Paz, Bolivia
[3]Department of Atmospheric and Oceanic Sciences, University of Maryland, College Park, MD, USA
[4]Institute of Physics, University of Sao Paulo, São Paulo, Brazil
[5]Federal University of Uberlândia, Agrarian Sciences Institute, Uberlândia, MG, Brazil
[6]Finnish Meteorological Institute, Helsinki, 00101, Finland
[7]Université Grenoble Alpes, CNRS, IRD, Grenoble INP, Institut des Géosciences de l'Environnement, Grenoble, France
[8]Leibniz Institute for Tropospheric Research, Permoserstr. 15, 04318 Leipzig, Germany
[9]Department of Environmental Science & Bolin Centre of Climate Research, Stockholm University, Stockholm 10691,Sweden

Correspondence: Victoria Sinclair (victoria.sinclair@helsinki.fi)

Abstract.

Observations of aerosol and trace gases in the remote troposphere are vital to quantify background concentrations and identify long term trends in atmospheric composition on large spatial scales. Measurements made at high altitude are often used to study free tropospheric air however such high-altitude sites can be influenced by boundary layer air masses. Thus,

5   accurate information on air mass origin and transport pathways to high altitude sites is required. Here we present a new method, based on the source-receptor relationship (SRR) obtained from backwards WRF-FLEXPART simulations and a k-means clustering approach, to identify source regions of air masses arriving at measurement sites. Our method is tailored to areas of complex terrain and to stations influenced by both local and long-range sources. We have applied this method to the Chacaltaya (CHC) GAW station (5240 m a.s.l.,16.35° S, 68.13° W) for the 6-month duration of the "Southern hemisphere

10   high altitude experiment on particle nucleation and growth" (SALTENA) to identify where sampled air masses originate and to quantify the influence of the boundary layer and the free troposphere. A key aspect of our method is that it is probabilistic and for each observation time, more than one air mass (cluster) can influence the station and the percentage influence of each air mass can be quantified. This is in contrast to binary methods, which label each observation time as influenced either by boundary layer or free troposphere air masses. We find that on average, 9% of the air sampled at CHC, at any given observation

15   time, has been in contact with the surface within 4 days prior to arriving at CHC, 24% of the air was located below 1.5 km above ground level and consequently, 76% of the measured air masses at CHC represent free tropospheric air. However, pure free-tropospheric influences are rare and often samples are concurrently influenced by both boundary-layer and free-tropospheric

air masses. A clear diurnal cycle is present with very few air masses that have been in contact with the surface being detected at night. The 6-month analysis also shows that the most dominant air mass (cluster) originates in the Amazon and is responsible

20 for 29% of the sampled air. Furthermore, short-range clusters (origins within 100 km of CHC) have high temporal frequency modulated by local meteorology driven by the diurnal cycle whereas the mid- and long-range clusters' (>200 km) variability occurs on timescales governed by synoptic-scale dynamics. To verify the reliability of our method, in-situ sulfate observations from CHC are combined with the SRR clusters to correctly identify the (pre-known) source of the sulfate: the Sabancaya volcano located 400 km northwest from the station.

1 Introduction

Traditionally, high-altitude measurement sites are used to study the remote atmosphere and the interactions between the free troposphere (FT) and the planetary boundary layer (PBL). These sites provide the opportunity to have long-term in situ observations of the free troposphere with high time resolution (Collaud Coen et al., 2018) as opposed to the short duration and inherent transient nature of airborne measurement campaigns. Observations of aerosol and trace gases in the FT are of great

30 scientific value in terms of understanding long-range transport and atmospheric chemistry, quantifying background concentrations, and observing long term changes in the composition of the atmosphere (Laj et al., 2009). Nevertheless, it is well known that high-altitude mountain sites can be influenced by boundary-layer air.

The planetary boundary layer is the lowest part of the atmosphere and is in direct contact with the Earth's surface. The majority of natural, and especially anthropogenic, aerosols and pollutants are emitted at the surface and thus directly released

35 in the PBL. The concentrations of gases and aerosols and their residence times depends on the structure of the PBL. Over land the PBL has a pronounced diurnal cycle with deep, well-mixed boundary layers typically observed during the day and shallow, stable boundary layers occurring at night. In complex terrain, thermally driven winds develop (e.g. slope and valley winds, De Wekker and Kossmann (2015)) during the day and can transport aerosols and pollutants from valley bottoms to high-altitude sites. Additionally, complex mountain meteorological processes such as orographic lifting can also transport PBL

40 air to high altitude.

Therefore, high-altitude sites can be influenced by local boundary-layer air and free tropospheric air that may, due to stronger upper-level winds, have undergone long-range transport. This means that observations must be carefully screened and analysed in synergy with many parameters to understand the dynamics and diurnal cycle of individual sites. Consequently, understanding the history of air masses sampled at mountain-top sites and related chemical composition is not an easy exercise. Since

45 the chemical and physical composition of the sampled air masses is, in general, inherently related to its path through the atmosphere (Fleming et al., 2012), having good methods to describe the history of the sampled air masses increases the value of measurements. Under these circumstances, the observations can then be treated as samples from different parts of the atmosphere both in the vertical (PBL vs FT) and the horizontal (short-range vs long-range contributions) domain.

Much of the analysis of ground-based atmospheric composition observations is accompanied by studies of what is known as

50 "climatological pathways" with the aim of mapping the probability of certain air masses reaching the station and identifying the sources and processes influencing certain types of air masses (Fleming et al., 2012). There are various ways of identifying air mass source regions. A popular and easy-to-apply approach to identify whether the measured air mass originates from the PBL or FT is to consider the presence of specific tracers such as Radon-222 (Griffiths et al., 2014) or the ratio of carbon monoxide to oxidized nitrogen species (CO / NO_y) (Herrmann et al., 2015). This tracer based approach has the limitation that

55 it cannot easily resolve horizontal source areas and only resolves two layers in the vertical. Alternative methods to identify source regions, which utilizing meteorological observations or numerical models, range from simply looking at the local wind direction to single trajectory analysis, Lagrangian dispersion models and even chemical transport models. The choice of method to link the atmospheric composition to the air mass history largely depends on where the expected sources are. If the sources

are predominately local, using local winds observations as a proxy for air mass history may suffice and this approach has been

60 taken by e.g. de Foy et al. (2008) and Salisbury et al. (2002). The major weakness of using in-situ wind measurements as an indicator of air mass history is that we cannot assume that the local measurements are representative of a larger region or of the synoptic-scale flow. This is particularly important for regions with complex topography in which the wind speed and direction at the surface may differ drastically to the wind aloft.

Alternatively, if sources are remote a low-resolution modelling approach would be more appropriate to identify air mass

65 history (Brattich et al. (2020); Sturm et al. (2013)). Single back trajectory models, often driven by low-resolution meteorological input data, have been widely used to gain insight into the sampled air mass history at high-altitude sites (Keresztesi et al. (2020); Brattich et al. (2020); Ghasemifard et al. (2019); Bolaño-Ortiz et al. (2019); Qie et al. (2018); Ou-Yang et al. (2017); Chauvigné et al. (2016); Brattich et al. (2015); Gratz et al. (2015); Ou-Yang et al. (2014); Putero et al. (2014); Tositti et al. (2013); Cheng et al. (2013), and Fleming et al. (2012) for earlier references). Advantages of this method are that it

70 is computationally efficient, easy to perform, and for individual case studies the output is simple to interpret. These studies primarily consider the horizontal, large-scale flow obtained by the single trajectory model. This is particularly problematic in high-altitude sites where the influence of the complex topography is known to have a clear effect on the interaction with the convective boundary layer (Serafin et al., 2018). Another disadvantage of single back trajectories, as shown by Stohl et al. (2002), is that they do not account for the filamentation and backward volume growth of the finite sampled air masses.

75 Lagrangian dispersion models have also been used to identify source areas at high-altitude sites (Ubl et al., 2017; Cécé et al., 2016; Lopez et al., 2015; Sturm et al., 2013; Brunner et al., 2012; Conen et al., 2012; Hirdman et al., 2010; de Foy et al., 2009). They differ from single back trajectory models by accounting for the deformation and stochastic dispersion of air masses due to turbulent mixing and convection (Stohl et al., 2002). A notable downside of dispersion models, particularly when applied to studies covering many months, is that they produces a large amount of output which is quite complicated to understand and

80 interpret. Therefore, additional post-processing and automated, objective analysis methods need to be applied to this output to extract the maximum amount of information while ensuring the resulting dataset is user-friendly. One way of post-processing dispersion model output is to perform a cluster analysis, where statistical methods can be used to differentiate different source regions.

Cluster analysis is a multivariate technique used to classify elements into groups in a way that maximizes the similarity (by

85 a predefined metric) within members of a group while also maximizing the dissimilarity across groups. Clustering has been extensively used in studies that aim to classify air mass history. In the case of single trajectory studies, the goal is to group trajectories into ensembles that follow a similar pathway (Kassomenos et al. (2010) and references therein). In dispersion models, clustering analysis has been applied both to classify the retroplume of the particles (e.g. Stohl et al. (2002)) and also to classify the regional footprint of the particles (e.g. Sturm et al. (2013); Paris et al. (2010)). However, these studies mostly

90 assume that high-altitude sites are also background sites, and therefore are mostly interested in the contribution of long-range transport sources. This means that the meteorological data used to drive the dispersion models is only required at low spatial resolution since the local sources are assumed to be negligible. However, there is still a lack of classification in the vertical

dimension and accountability for the influence of short- and long-range transport simultaneously which is of special relevance for locations where short-range sources are equally as relevant as more distant sources.

95   One such high-altitude site is the GAW (Global Atmospheric Watch) Chacaltaya (CHC) atmospheric research station (16.35°S, 68.13°W) located at $5240\,\mathrm{m}$ a.s.l. and 20 km from the metropolitan area of La Paz / El Alto but ∼1.6 km higher in altitude than the centre of La Paz. For a detailed description of the site see Chauvigné et al. (2019) and Wiedensohler et al. (2018). Measurements of reactive and greenhouse gases as well as aerosol optical, chemical and physical properties are routinely monitored at the station following the GAW recommended procedures (Laj et al., 2020). At this station, in the con-
100   text of the SALTENA (Southern hemisphere high altitude experiment on particle nucleation and growth) campaign (Bianchi et al., 2021, in review), state-of-the-art instruments that measure aerosol chemical and physical properties were deployed to complement on-going long-term observations. The intensive measurements took place between December 2017 and June 2018 (covering both wet and dry seasons). The unique location of the station in the under-sampled southern hemisphere enabled us to study a mixture of pristine air masses from the Amazon Basin loaded with biogenic emissions, regional background air masses
105   from the Altiplano perturbed by volcanic activity, and, marine air masses from the Pacific Ocean. In addition, strong anthropogenic influence from the La Paz / El Alto metropolitan area was sampled. This wide range of potential source areas, along with complex mountain meteorology, and state-of-the-art, highly detailed observations of the physical and chemical properties of aerosol and trace gases, means that a comprehensive meteorological analysis, beyond what is the typical performed for aerosol measurement campaigns, is required.

110   The overall objective of this study is to develop a new method to identify air mass source regions which is valid for high-altitude stations that are influenced both by local and long-range sources and where the vertical classification of sources is as relevant as horizontal segregation and then to apply this method to CHC. An overview of the method is given in Fig. 1. The first aim of this study is, therefore, to use a regional meteorological model (Weather Research and Forecasting model - WRF) in combination with a Lagrangian dispersion model (FLEXible PARTicle dispersion model - FLEXPART) to create a high-
115   resolution data set of source areas for CHC at hourly resolution (steps 1 and 2 in Fig. 1). The second aim is to develop a new method, based on cluster analysis, to transform the complex, very large dataset into a user-friendly dataset of air mass source regions (steps 3 to 7 in Fig. 1). The third aim is to document the characteristics of the identified source areas (clusters) which will enable the dataset produced here to be applied in forthcoming studies on the chemical composition measurements made during the unique high-altitude SALTENA campaign. The fourth and final aim is to demonstrate the strength and simplicity of
120   the classification results from our method which we do by using them to confirm a well-known source of sulfate emissions that were measured at the CHC station.

The remainder of this study is structured as follows. In Section 2, the meteorological model and the Lagrangian dispersion model are described. The newly developed clustering method is described in Section 3. Additional diagnostics are presented in Section 4. The relative contribution of the PBL and the FT to CHC are described in Section 5.1. The characteristics of the
125   identified source regions are discussed in Sections 5.2 and 5.3. An example indicating that the method works well is shown in Section 5.4. A Discussion on the results and recommendations are presented in Section 6. Finally, the conclusions are presented in Section 7.
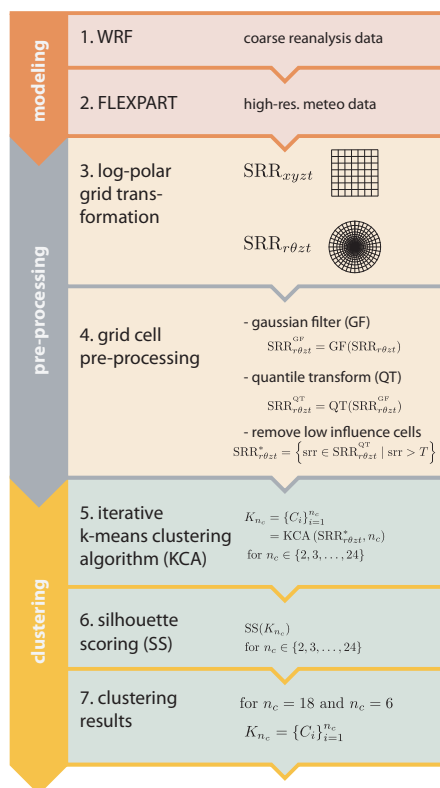
Figure 1. Flowchart describing the method's steps. The steps are divided into three groups: modelling, pre-processing and clustering. SRR refers to the source-receptor relationship (explained more in Section 3) and T is the threshold used for the removal of low influence cells. K_{n_c} is the set containing n_c number of clusters C_i. Each cluster C_i is populated with grid cells that have a similar temporal influence over CHC.

2   High resolution meteorological modelling and backward dispersion simulations

To generate a high-resolution, gridded dataset of meteorological variables that can be used to drive a Lagrangian dispersion model, we used the Weather Research and Forecasting (WRF) model version 3.7. WRF is a state-of-the-art, non-hydrostatic, regional numerical weather prediction model that is used operationally and for research. Here we perform one 6-month long continuous simulation starting on 2017-12-06 and ending on 2018-05-31. Four nested domains are used and their locations are shown in Fig. 2a and c. The outermost and largest domain has a grid spacing of $38\,\mathrm{km}$ whereas the innermost and smallest domain has a grid spacing of $1\,\mathrm{km}$. One-way nesting is used: the outer domain provides boundary data for the inner nest but the inner nest does not provide any feedback to the outer domain. To ensure the long simulation remains close to reality throughout the 6-month period, the outer domain is nudged (i.e. analysis nudging) to the reanalysis fields every 6 hours.

The initial and boundary conditions were taken from the NCEP Climate Forecast System reanalysis Version 2 (Saha et al., 2013) with a temporal resolution of $6\,\mathrm{hours}$, a nearly $100\,\mathrm{km}$ horizontal resolution and 64 sigma-pressure hybrid layers. The
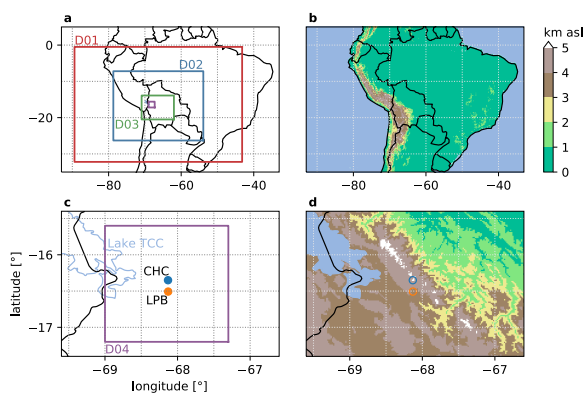
Figure 2. Overview of the studied region and the WRF model domains. Panel a) shows the location of the 4 nested domains (D01, D02, D03 and DO4), b) shows the topography in kilometers above sea level (km a.s.l.) of the whole domain, c) is a zoomed in version of a) and also shows the location of the Chacaltaya station (CHC, $5.2\,\mathrm{km}$ a.s.l. blue dot), La Paz / El Alto metropolitan area (LPB, $3.6\,\mathrm{km}$ a.s.l., orange dot) and Lake Titicaca (TCC, $3.9\,\mathrm{km}$ a.s.l. blue outline). Panel d) shows the topography (km a.s.l.) in the area closest to CHC.

model topography was obtained from the Global multi-resolution terrain elevation data model (GMTED2010) (Danielson
140 and Gesch, 2011) with a resolution of $1\,\mathrm{km}$. We use the following parameterizations: microphysics is parameterized by the
Goddard Scheme; cumulus convection is parameterized by the Grell–Freitas Ensemble Scheme in D01 and D02 and no cumulus
parameterization is used for D03 and D04; the short and long-wave radiation is parameterized by the New Goddard Shortwave
and Longwave Schemes; the Planetary Boundary Layer (PBL) Physics are represented by the Mellor–Yamada–Janjić Scheme
(MYJ); and the land surface model is the Unified Noah Land Surface Model.

145 The surface temperature of Lake Titicaca was manually prescribed to monthly means obtained from Pillco Zolá et al. (2019)
since values prescribed by WRF were unrealistically low. This is most likely due to the height of the lake (around $3.8\,\mathrm{km}$ a.s.l.)
and the assumptions made by WRF when interpolating surface temperature of lakes from adjacent sea surface temperature.

There are some limitations to the WRF simulation, primarily related to the complexity of the terrain. The 1-km resolution
inevitably smooths the topography in comparison to reality which can affect the slope angles and furthermore affect the sim-
150 ulated thermally-driven winds. However, wind, temperature, and precipitation comparisons with in situ observations at CHC
(not shown) show reasonable agreement.

The FLEXible PARTicle dispersion model (FLEXPART) is a Lagrangian transport and dispersion model which can be
used for both forward and backward simulations. We used version FLEXPART-WRF_v3.3.2 (Brioude et al., 2013) to perform
backward simulations and thus to determine the source regions of air masses arriving at CHC. The FLEXPART simulations
155 were driven using the meteorological output from the 6-month WRF simulation (described in Section 2). Output from all
four of the WRF output domains was used and this was available at a temporal resolution of 15 minutes. This high temporal
resolution is a clear advantage over using reanalysis data which at best is only available once per hour.

In the FLEXPART simulations, we continuously release $20\,000\,\mathrm{particles/hour}$ from CHC from 2017-12-06 00 until 2018-07-01 00 and compute their back trajectories for $4\,\mathrm{days}$. The particles, passive air tracers, are released in a 10-m deep layer which extends from $0\,\mathrm{m}$ to $10\,\mathrm{m}$ a.g.l. and over a $2\,\mathrm{km} \times 2\,\mathrm{km}$ square centred around CHC.

When FLEXPART is run in backward mode, it calculates the emission sensitivity response function, also referred to as the source-receptor relationship (SRR), on a user-specified three-dimensional longitude-latitude-height grid. The output of FLEXPART can be in different units and here we configure the model so that the output (SRR) is in seconds.

FLEXPART also permits the user to specify two nested output domains with the inner domain, closest to the release site (i.e. the receptor), having a higher resolution than the outer domain. We make use of this functionality and specify the first FLEXPART output domain to have the same geographic extent as D03 in the WRF simulation but the spatial resolution ($1\,\mathrm{km}$) as D04 in the WRF simulation. Our second, outer FLEXPART output domain covers the same region as D01 in the WRF simulation and has the same resolution ($38\,\mathrm{km}$) as D02 in WRF. In the vertical direction, we specify the FLEXPART output grid to have 30 uniform levels each 500 m deep extending from the surface to 15 km a.g.l.

FLEXPART also contains options, controlled via namelist options, for how turbulence and convection are included in the simulations. We take the values of the PBL height, surface sensible heat flux, and the friction velocity directly from the WRF simulation (SCF_OPTION=1). Turbulence is parameterized using the Hanna scheme (Hanna, 1984) as used in FLEXPART-ECMWF/GFS (TURB_OPTION=1) and we assume skewed rather than Gaussian turbulence in the convective boundary layer (CBL=1). Deep convection is also parameterized (LCONVECTION=1).

3 Dispersion simulation output pre-processing and clustering

Here we describe the core of our new method, namely steps 3 to 7 in the methods diagram (Fig. 1). The output from FLEXPART is the source-receptor relationship (SRR) which is related to the particles' residence time in the output grid cells. The grid cells are defined by a regular longitude (x), latitude (y), and height (z) grid SRR_{xyzt} where the t dimension is related to the particle's release time. The output is on 2 nested grids, which means we obtain high-resolution information for regions that are near the receptor and lower resolution information for more distant locations. The rationale is that near the receptor, the higher resolution provides better detail on potential high-influence sources, for example, in our case, the $20\,\mathrm{km}$-away nearby metropolitan area of La Paz / El Alto. Furthermore, far away, a low-resolution grid cell suffices since localized potential source influences are diffused. However, the interface region between these two domains is potentially under-sampled on one side and over-sampled on the other side. Also, a priori information on where to locate the interface between these 2 domains is unlikely to be available.

To overcome these limitations, in this study we propose the use of a log-polar transform (Sarvaiya et al., 2012) of the coordinate system (Fig. 1 step 3). The new log-polar grid has grid cells which gradually increase in size as the distance from the receptor increases with no sudden step changes to the resolution. Since we are in the tropics, we can safely use an equirectangular projection to a Cartesian coordinate system defined by longitude (lon), latitude (lat) and height above ground

level (z). Any point (lat, lon, z) can be represented in polar cylindrical coordinates and is given by

190
$$r = \sqrt{(\text{lon} - \text{lon}_c)^2 + (\text{lat} - \text{lat}_c)^2}$$

$$\theta = \tan^{-1} \frac{\text{lon} - \text{lon}_c}{\text{lat} - \text{lat}_c}$$

$$z = z$$

where r is the radial distance to the receptor location (lon_c, lat_c) and θ is the clockwise angle starting north from the receptor. Notice that r is the Euclidean distance of lat and lon. The relation between r and the geodesic distance d in km is given by the
195   approximation

$$d\,[\text{km}] = 108.6 \frac{\text{km}}{^\circ}\, r\,(\pm 3\,\%)$$

and is valid for the whole region covered by the WRF D01 domain. The radial boundaries of the log-polar grid are separated by a distance $\Delta\theta = 10°$. The radial length Δr of the log-polar cells is $\Delta r = r_{i+1} - r_i$ where $r_{i+1} = r_i e^a$. The value of $a = 0.18$ is chosen so that the log-polar cells approximate a square with sides $\Delta r \approx \Delta\theta$. The ring radii of the log-polar grid are determined
200   by starting with a initial ring r_0 of radius $0.08°$ ($\approx 8.7\,\text{km}$). The choice of the value for r_0 should be large enough to allow the first ring of radial cells to have an area larger than the grid cells in the highest resolution output from FLEXPART so that at least one original FLEXPART output grid falls on each radial cell. The following 30 ring radii are obtained iteratively using $r_{i+1} = r_i e^a$. For each log-polar grid cell, the SRR is obtained by adding the SRR values of the rectangular grid cells whose centre of mass is contained within the log-polar grid cell. As a result of the grid transformation, we are left with $33\,480$
205   log-polar grid cells defined by 36 θ planes, 31 r cylinders and 30 z levels: $\text{SRR}_{r\theta z t}$.

In order to perform a cluster analysis on this dataset, we need to define the elements that will be clustered/grouped, the number of clusters, and the features of each element used to determine the group identification. The 33480 log-polar grid cells are taken to be the elements to be clustered. The element's features are the SRR intensities at each release time (although called the release time, it is indeed the arrival time at the destination i.e. the receptor). The particles were released at hourly intervals
210   starting on 2017-12-06 and ending on 2018-07-01, therefore, there are 4248 features for each element.

It is common practice to pre-process the elements before applying the clustering algorithm. In our case, we apply three pre-processing procedures (Fig. 1 step 4): Gaussian filter smoothing, quantile transform and filtering out non-relevant elements. In order to smooth the dataset, we order the elements in a 4-dimensional array. Three of the dimensions are defined by the previously described cylindrical log-polar grid and the fourth one is the release time. The smoothing is obtained by applying a
215   multidimensional Gaussian filter on all 4 dimensions of the array.

We are interested in grouping elements depending on their variation over time rather than their absolute values, therefore we need to normalize the dataset. We use scikit-learn's (Pedregosa et al., 2011) quantile transform function to normalize our elements to a uniform distribution. This procedure has the advantage of being robust to outliers and also performs quite well with sparse arrays or semi-sparse arrays like ours. After this function is applied, the distribution of each element resembles a
220   uniform distribution with a value range from 0 to 1.

Finally, we filter out elements whose SRR values are either zero or have so little influence that including them adds computational burden to the clustering algorithm and does not improve the results.

In order to decide which elements are not beneficial based on the above definition, we define a threshold T in the following way. First, we sum all the SRR values for each of the elements over the time period. Then, we sort the element based on their total SRR value and compute the cumulative SRR values. Finally, we split this dataset at the point where 85 % of the total SRR value is reached and discard the remaining 15 %. This procedure leaves us with 8580 cells out of the total 33 480 in the grid. Out of the excluded cells, 24.8 % had a zero SRR total value and the total median for the non-zero left out cells is 2060 s. The total median for the included 8580 cells is 112 190 s.

The next step (Fig. 1 step 5) consists of clustering the pre-processed $SRR_{r\theta zt}$, i.e., the goal is to divide the $SRR_{r\theta zt}$ matrix into a number n_c of groups (clusters), whose cells have a similar temporal evolution regarding their SRR values. The procedure is similar to the one used by Herrmann et al. (2015) and references there-in with a few notable differences outlined below. The overall objective is to cluster 3D grid cells in the domain based on the SRR contribution that they have over time as opposed to clustering the spatial patterns (snapshots). We use the 'k means' (Lloyd, 1982) clustering algorithm due to its generalized use, speed and adequate performance with a large number of elements and medium size number of clusters.

However, the 'k means' algorithm does not automatically select the number of clusters and there is not a right answer for the number of clusters. Too few clusters (e.g. 2) means that no meaningful information is obtained whereas too many clusters (e.g. 100) is impractical and risks overfitting. The optimal number of clusters is usually determined by trying a variety of options and calculating quantitative measures of how similar an element is to its own cluster compared to other clusters. Here we
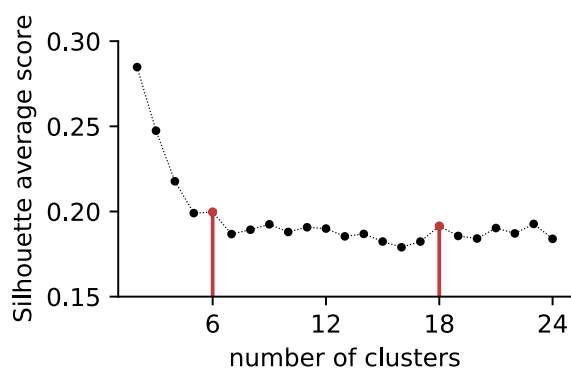
Figure 3. Average Silhouette score for the iterative 'k means' clustering algorithm as a function of number of clusters (from K=2 to 24). Red vertical bars and points show the selected number (6 and 18) of clusters.

try 2 to 24 clusters and use the silhouette method (Rousseeuw, 1987), which ranges from -1 to +1, to determine the optimal number (Fig. 1 step 5). Large positive values of the silhouette score indicate that the element is well matched within its clusters and poorly matched to other clusters while a low or negative value imply possible miss-classification. The overall silhouette score is obtained by averaging over the scores for each individual element and in our case we do a weighted average based

on the total SRR of each element, so that cells with a high density have a bigger influence on the overall score. The resulting silhouette scores for 2 to 24 clusters are shown in Fig. 3 and we selected the cluster number where a localised maximum

245 occurs. Two options exist: 6 clusters or 18 clusters (Fig. 1 step 7). A second approach to select the number of clusters, which is complementary to and can be used together with the silhouette score, is to consider the number of clusters needed to address the scientific question. In our case, we are interested in the air mass footprint not only in terms of their direction from CHC but also the distance from the station and the vertical heights. Therefore a satisfactory solution would have approximately 16 clusters: 4 directions x 2 height levels x 2 distance ranges. As this is in rough agreement with the 18 clusters we find using the

250 silhouette score, we select 18 clusters. This is however quite a large number and for some applications may be too detailed. Therefore, in addition we also perform the analysis for 6 clusters. To ensure the 6 clusters can be directly related to the set of 18 clusters, we first create the 18 clusters and then perform a second round of clustering (starting from the 18 clusters) to obtain the 6 clusters. The 6 clusters are subsequently referred to as the 6 main pathways.

4 Additional diagnostics and data

255 In order to understand and categorize air masses that have been in contact with either the surface (where the emissions occur) or the PBL (where surface emissions are well mixed), we define the surface SRR percentage influence:

$$S_{\mathrm{SRR}} = \frac{\mathrm{SRR}_{\mathrm{surface}}}{\mathrm{SRR}_{\mathrm{total}}} \times 100 \tag{1}$$

and the pseudo boundary layer percentage influence :

$$\mathrm{PBL}^*_{\mathrm{SRR}} = \frac{\mathrm{SRR}_{<1.5\,\mathrm{km}}}{\mathrm{SRR}_{\mathrm{total}}} \times 100. \tag{2}$$

260 Using the concept of a pseudo boundary layer, which has a constant depth of 1.5 km (i.e we select the first 3 levels of the FLEXPART output ($<1.5\,\mathrm{km}$), means that we neglect the diurnal variation in the PBL height. We make this approximation as due to our computational procedure, the specific depth of the PBL is lost when transforming the SRR output into log-polar coordinates. The choice of 1.5 km is motivated by previous studies that have quantified PBL depth in nearby regions. For example, Carneiro and Fisch (2020) analysed radiosonde and remote sensing data from the GoAmazon project (Martin et al.,

265 2016) and show that the typical minimum PBL height is 250 m and the deepest PBLs occur during daytime in the dry season and are 1.5 km deep on average. The global study by von Engeln and Teixeira (2013), based on reanalysis data, shows that PBL heights are somewhat deeper near CHC than in the Amazon but PBL heights typically range between 500 m and 1.5 km. Therefore, the real PBL will usually be similar in depth or shallower than our value of 1.5 km. This means using the pseudo BL depth will likely over-estimate the influence of PBL air masses.

270 When referring to the SRR percentage influence of a cluster, for simplicity we use

$$\mathrm{SRR}[\%] = \frac{\mathrm{SRR}}{\mathrm{SRR}_{\mathrm{total}}} \times 100 \tag{3}$$

where the SRR_{total} is equal to the theoretical total residence time of the simulation expressed in seconds ($4\,\mathrm{days} = 345\,600\,\mathrm{seconds}$). This means that for every time step of the FLEXPART simulation, the sum of all the SRR cluster values adds up to SRR_{total}. In practice this theoretical value is not always achieved since inevitably a very small fraction of the particles leave the outer domain (D01) before the end of the 4-day simulation. This also implies that for some time periods, the sum of all the cluster might not add up to 100%.

In order to understand diurnal cycles in the SRR from different vertical levels and regional sources, we generated detrended diurnal cycle timeseries by decomposing the SRR raw signal into a trend (SRR_{trend}) and a diurnal cycle ($SRR_{diurnal\ cycle}$) in the following fashion:

$$SRR_{diurnal\ cycle}(t) = SRR_{raw\ signal}(t) - SRR_{trend}(t). \qquad (4)$$

The SRR_{trend} is obtained by applying a running mean with a window of $24\,\mathrm{hours}$ to the raw signal.

Furthermore, when analysing the diurnal variation of the SRR for each of the 18 clusters we first objectively identify the clusters that present a diurnal cycle by transforming their time series into the frequency domain by applying a fast Fourier transform and plotting the resulting frequency spectrum (Fig. S3). In this domain, we visually identify those clusters that present a local peak around the frequency $1\,\mathrm{day}^{-1}$. Manual inspection of the clusters' timeseries indicated that not every day included in each cluster identified by the fast Fourier transform process to have a peak at $1\,\mathrm{day}^{-1}$ had a pronounced diurnal cycle. Therefore, the clusters are then subjected to a dynamic time warp (DTW) grouping procedure (Tavenard et al., 2020) that classifies the days with and without a diurnal peak i.e. influence from the boundary layer. The results of this classification are shown in Figs. S4 to S7.

In order to describe the land-use characteristics of the geographical areas that the resulting clusters and main pathways originate from, we make use of the World Wildlife Fund (WWF) terrestrial ecoregions (TER, Dinerstein et al., 2017), WWF marine ecoregions (MER, Spalding et al., 2007) classification scheme, and two extra regions (EXR), namely the Lake Titicaca and the Southern Atlantic ocean regions, that we add to fulfil the domain. In total, 37 different ecoregions exist in the area covered by the largest WRF model domain. The advantage of using these regions is that they are well defined and described in the literature, do not depend on arbitrary political borders and are defined considering regions that have similar ecosystems and therefore similar potential emission signatures. Furthermore, the regions are also nested within biomes that provide a more general picture. Within our outermost domain (D01), 13 different biomes are present.

Sulfate mass concentrations measured at the CHC station from March to May 2018 are used in this study to illustrate an example application of our clustering methodology. The aim of this example is to show that our new method can identify the main source of the high sulfate concentrations measured at CHC. A very likely source of this is degassing from nearby volcanoes which was observed at the same time. The sulfate dataset is obtained using the quadrupole aerosol chemical speciation monitor (Q-ACSM, Aerodyne Research Inc.) which is able to routinely characterize non-refractory submicron aerosol species such as organics, nitrate, sulfate, ammonium, and chloride (Ng et al., 2011). Because of the low atmospheric pressure, a 130 μm diameter critical orifice was used in order to retain the normal sample mass flow rate (Fröhlich et al., 2013). In addition, inlet flow and mass calibration (using ammonium nitrate and ammonium sulfate as standards) were accomplished to guaran-

tee optimal instrumental performance and mass quantification. The instrument's time resolution was 30 minutes. This sulfate timeseries was then correlated to the SRR timeseries of both the 18 clusters and the 6 main pathways.

5 Results

We now present the results. In section 5.1 we quantify the contribution of the surface, "pseudo" PBL, and FT sources to the
310 air masses measured at CHC. In section 5.2 the characteristics of the 6 main pathways are presented before the more detailed analysis of the 18 clusters is shown in in section 5.3. Finally, in section 5.4 we show one example of how our clustering results can be combined with measurements to identify source areas.

5.1 Relative contribution of PBL and FT to air sampled at CHC
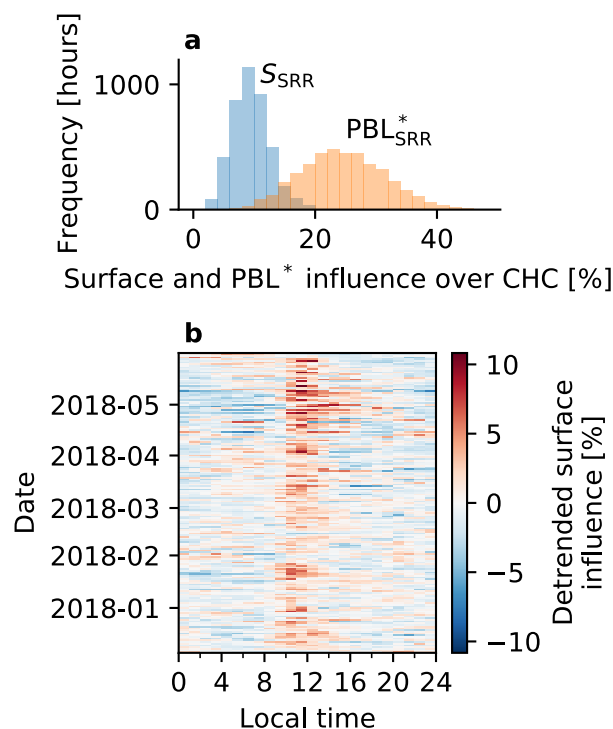
Figure 4. Panel a) is a frequency histogram showing the percentage of air sampled at CHC during each modelled hour that was in contact with the surface (S_{SRR}, Eq. (1), blue) or located within the pseudo PBL (PBL^*_{SRR}, Eq. (2), orange). Panel b) shows the detrended surface influence for every day of the modelled period as a function of local time.

As stated in the introduction, high-altitude mountain-top sites are often used in an attempt to sample free tropospheric air. In this section, we use the SRR output from FLEXPART to determine the relative influence of PBL air and FT air on the total

sampled air mass at CHC over the 6-month period and the diurnal pattern of this influence. Previous methods, for example, those based on specific tracers (Sturm et al. (2013)), are binary in nature as they label each observation time as influenced either by PBL or FT air masses. In contrast, our method is probabilistic and for each observation time determines the percentage of the sampled air mass influenced by the PBL or FT. Figure 4a shows frequency histograms of the percentage influence of the "surface" and of a pseudo PBL (see Eqs. (1) and (2) for definitions) for each hour. There is a strong linear correlation between the surface and PBL influence; the best fit regression is given by

$$S_{SRR} = 0.42 \times \mathrm{PBL}^{*}_{SSR} - 0.76$$

and the r-squared value is 0.9 .

315   The median influence of the surface is 9% meaning that on average 9% of the air sampled at CHC has been in contact with the surface in the last 4 days. In terms of the pseudo PBL, on average 24% of the sampled air masses represent PBL air. Indirectly this means that approximately 76% of the air sampled at CHC can be considered representative of the FT. Note that this does not mean 76% of observation times are representative of the FT; it should be interpreted as, on average, at any given time ∼24% of the measured air mass represents the PBL and the remainder the FT. This is a key strength of our method;
320   it can determine at any given time what percentage of the sampled air mass arrives from different locations. An additional interpretation of the results shown in Fig. 4a is to consider the percentage of time when there is no influence (0 % on the x-axis) from the surface or pseudo BL i.e. where there is purely free tropospheric air masses. This situation is never detected which indicates CHC is rarely representative of purely FT air. However, this is partly an artefact of the method employed here as all particles are forced to arrive at the surface (10 m a.g.l.) at the station.

325   Figure 4b shows the diurnal cycle of the detrended surface influence for each day of the 6-month campaign. The largest positive values, indicative of a large surface influence, occur during daytime. The peak emerges at 10 am local time almost every day. The duration of the high surface influence increases throughout the campaign with higher values of surface influence extending later in the afternoon in April and May. This gradual increase in the surface influence during the campaign might be explained by the transition towards the dry season when clear-sky conditions become more frequent increasing insolation
330   periods which, in turn, favour deep well mixed PBL structures.

5.2   The six main pathways

In this section we describe the results from clustering the SRR log-polar cells into 6 groups. We call these 6 groups the main pathways (PW) since they tend to start near the station and reach far away from it as opposed to the 18 clusters that occupy more localised regions. Furthermore, we label each cluster based on their 'clock direction' from CHC and append them with
335   the acronym PW to distinguish them from the 18 clusters. For example, cluster label '03_PW' refers to the cluster whose centroid position is located east from the station.

    In Fig. 5a and b we show the 6 main pathways (PW) along with the 18 cluster centroids. The pathways are the shaded coloured regions and contain 2 to 4 clusters from the 18-cluster grouping. The 03_PW is located geographically in the lowlands

to the east of the station occupying the biomes 1, 2, 6, and 9 which in general are tropical forests and grasslands (Fig. 5e and
340   f).

Cluster 05_PW originates from the south of the station in the altiplanic (montane grass and shrubland, biome 4) and lowland
(subtropical dry broadleaf, biome 6) regions. Horizontally, it follows the Altiplano plateau and its eastern slopes to the station.
The cluster 07_PW comes from the south-west and most of its area is located above the Pacific Ocean/coast (biomes 3, 5, 7
and 8). Cluster 09_PW comes from a relatively short distance and it occupies the altiplanic biome 4, the Titicaca lake (13)
345   and two Pacific biomes (5, 8). The cluster 11_PW comes from the north and north-west. It occupies the lowlands, Altiplano
and the Pacific coast i.e. biomes 1, 4, 11. As it gets closer to CHC, cluster 11_PW is located higher than 12_PW and thus
goes above 12_PW. Finally, 12_PW comes from the lowlands north of the station and is contained within biome 1 (tropical
broadleaf forest). In Fig. 6a to d, we respectively describe the pathways' centroids in terms of their height above ground;
height above sea level; surface influence (S_{SRR}, Eq. (1)); and SRR percentage influence. In general, the farther the centroid
350   is from CHC, the higher above ground level its centroid location is. The same pattern is observed for their height above sea
level, however, if the location of the centroid is not too far away and above a location where the ground height is considerably
lower than CHC, then the centroid location is located below the linear trend, e.g. 03_PW and 12_PW. In terms of their S_{SRR},
a decreasing trend is observed, in other words, the farther the centroid, the lower the influence from the surface. Therefore,
12_PW is highly influenced by its contact with ground (62 %, Fig. S1) while 07_PW is almost unperturbed by the surface
355   (8 %). Finally, the SRR influence of each pathway seems to be uncorrelated with the distance from CHC. Pathway 03_PW has
the highest influence over CHC with a share of 29 %. This is in agreement with previous studies at CHC (Chauvigné et al.,
2019) where air masses from the Amazon were identified as the major contributors during the wet season DJFM (our modelling
period involves DJFMAM).

Finally, in Fig. 7, we show the temporal influence of each of the pathways. We quantify the influence in percentage of each
360   pathway by dividing the pathway's SRR values by the theoretical total residence time of the simulation, 96 hours × 3600
seconds. Note that the sum of the influence for all 6 pathways shown in Fig. 7 does not always sum to exactly 100 which is
due to particles leaving the domain. One of the strengths of our method is that at any given time, we can identify more that one
cluster influencing the station.

A clear change in the influence pattern at the beginning of May is seen: On one hand, the influence of pathways 03_PW and
365   05_PW and to a lesser extent 12_PW becomes almost negligible. On the other hand, pathways 07_PW and 08_PW increase
their influence. This is consistent with Chauvigné et al. (2019) where it was shown that during the wet season (DJFMA) air
masses from the lowlands and the east-southeast tend to have a bigger influence on CHC. In our case, 12_PW and 03_PW are
clearly lowland pathways while 05_PW has a mixture of Altiplano and lowland influence that, nonetheless, comes from the
southwest and therefore is mostly favoured during the wet season. The pathway 11_PW does not present a clear change during
370   the 6 month period. Finally, visual inspection shows that the influence of each pathway happens in 1-to-2 week timespans. We
will further develop this point when focusing on the detailed 18 clusters (Section 5.3).
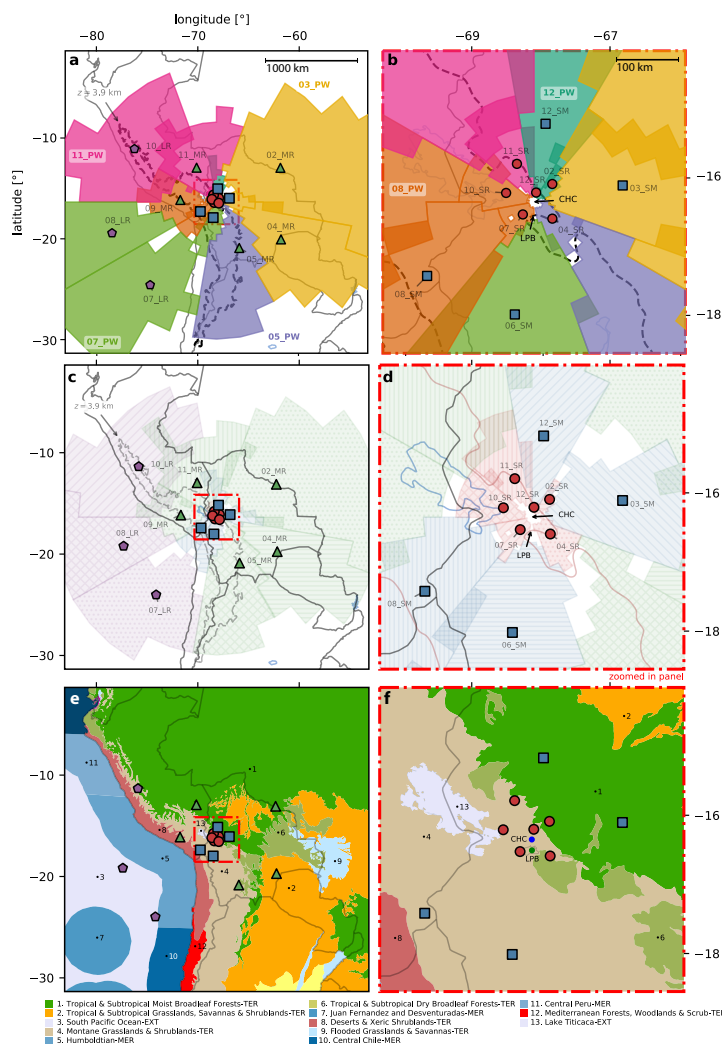
Figure 5. The shading in panels a) and b) show the horizontal extent of the 6 main pathways (PW). For easier visualization we only shade grid cells which have SRR values in the top 80% for each cluster. Each cluster centroid is marked with a disk (short range), square (short-medium range), triangle (medium range), or pentagon (long range) locator. The colour of each region is related to the main pathway they belong to: 03_PW (yellow), 05_PW (purple), 07_PW (green), 08_PW (orange), 11_PW (pink), and 12_PW (teal). The dashed black line corresponds to a height of 3.9 km a.s.l. and encircles the Altiplano plateau. Panel b) corresponds to the region inside the red rectangle in panel a). Panels c) and d) show the horizontal location of the 18 clusters. The colour is related to their distance range from CHC and the hatch distinguishes each cluster independently. Panel d) corresponds to the region inside the red rectangle in panel c). The colours, both for the centroid markers and the shaded areas, differentiate the distance range from CHC (SR, SM, MR, LR). The hatch patterns distinguish the area of each of the 18 clusters. Panels e) and f) show the biomes as described in Section 4 and are shown to compare the area of the clusters and pathways to the underlying biome. The resulting area overlaps between each cluster and biome are shown in Fig. 8
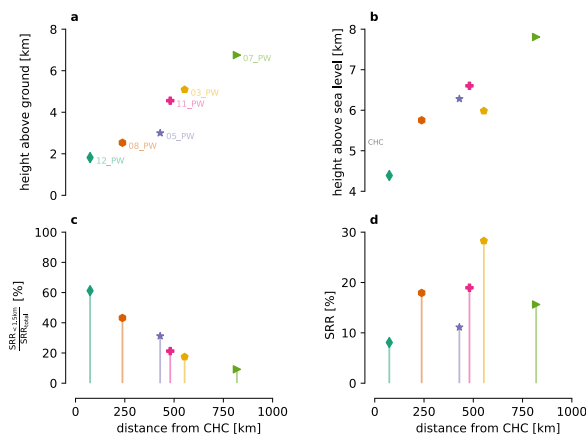
Figure 6. Centroid properties for each of the main pathways (PW). Panel a) shows the median height above ground level of each cluster while panel b) shows the median height above sea level. Panel c) shows the ratio between the SRR values that are below $1.5\,\mathrm{km}$ above ground level and the total SRR value for each cluster. Panel d) shows the mean SRR percentage for each PW. Quantitative numbers are presented in Fig. S1.
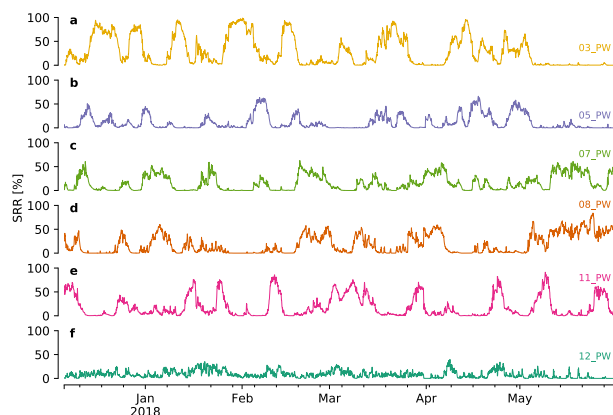
Figure 7. Time series of the SRR [%] influence for each of the 6 main pathways (PW) from 2017-12-06 until 2018-05-31.

5.3 The 18 clusters

In the previous section, we provided a general picture of the air masses that influence CHC by using the 6 pathways. However, for an in-depth description, we now focus on the more detailed 18 clusters. We have subdivided these 18 clusters into 4 sub-

375 groups based on their horizontal distance from CHC: short range (SR), short-medium range (SM), medium range (MR) and long range (LR) for distances ranging from 0 to $100\,\mathrm{km}$, 100 to $300\,\mathrm{km}$, 300 to $800\,\mathrm{km}$ and $>800\,\mathrm{km}$ respectively. Further-more, we have labelled each cluster based on their distance range along with their 'clock direction' from CHC. For example, cluster label '09_MR' refers to the cluster whose centroid position is located west from the station at a distance between

300 and 800 km. As done before, we first describe the clusters in terms of their horizontal location, then their vertical centroid properties and finally their temporal evolution. In Fig. 5c and d, we show the clusters geographical location along with a reference to the biomes that they occupy in space (Fig. 5e and f).

To further quantify the superposition of clusters and biomes presented in Fig. 5e and f, in Fig. 8 we show a heat map of the percentage of area that each cluster and biome share. For example, the geographical area of cluster 02_MR is split between the Tropical & Subtropical Moist Broadleaf Forests (biome 1, 56%), the Tropical & Subtropical Grasslands, Savannas & Shrublands (biome 2, 28%), the Tropical & Subtropical Dry Broadleaf Forests (biome 6, 10%), and the Flooded Grasslands & Savannas (biome 9, 6%). In general, there is a clear split between clusters located northeast (i.e. clock direction of 11 to 05) and southwest (clock direction of 05 to 11) from the station. The first are located in the generally more tropical and humid lowland biomes (1, 2, 6, 9) while the latter are located in drier altiplanic biomes (4, 13) and pacific biomes (3, 5, 7, 8, 10, 11, 8).
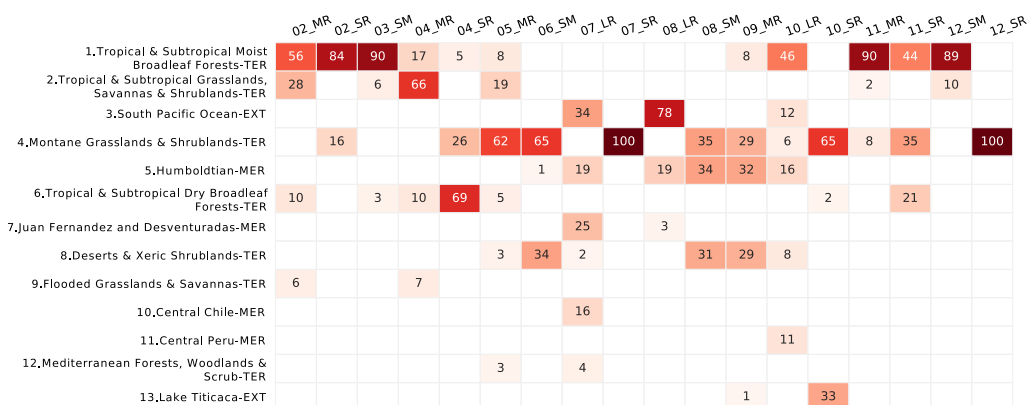
Figure 8. Heat map showing the biome and land cover characteristics associated with each of the 18 clusters. The percentages values indicate the percentage of air in each cluster that travels over each different biome. The darker the colour, the stronger the influence from the corresponding biome. For a description of the biomes see Section 4. Due to rounding errors and the use of integers, some columns add to 99 rather than 100.

In Fig. 9a to d, we respectively describe the clusters' centroids in terms of their height above ground; height above sea level; surface influence (S_{SRR}, Eq. (1)); and SRR percentage influence. In general, they follow the same patterns that we described for the 6 pathways with the exception that there are more clusters below the CHC station's height a.s.l. (Fig. 9b). These clusters, namely 02_SR, 12_SM and 03_SM are located close to CHC in the lowlands located north and northeast from the station.

Specifically, all of the short-range clusters' centroids are below 2 km a.g.l. and below 5.2 km a.s.l. (Fig. 9a and b) which is below the CHC station height (5.2 km a.s.l.). Furthermore, these clusters are in close contact with the surface since their S_{SRR} (Eq. (1)) is greater than 50 % (Fig. 9c).
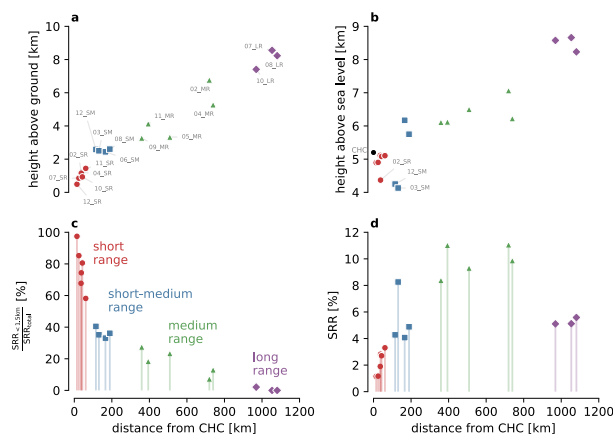
Figure 9. Cluster-median properties for each cluster. The horizontal axis for all panels represents the radial distance from CHC. Panel a) shows the median height above ground level of each cluster while panel b) shows the height above sea level. Panel c) shows the median ratio between the SRR values that are below $1.5\,\mathrm{km}$ above ground level and the total SRR value for each cluster. Panel d) shows the median SRR percentage for each cluster.

The short-medium range clusters' centroids are between $2.4\,\mathrm{km}$ and $2.6\,\mathrm{km}$ a.g.l. However, their height a.s.l. varies from 4.1 to $6.1\,\mathrm{km}$. This is due to clusters coming from both the Altiplano to the southwest and the lowlands to the northeast of the station. In general, only one third of these clusters' air masses are below $1.5\,\mathrm{km}$ (S_{SRR}) and thus these clusters include notably

400   more influence from the free troposphere than the SR clusters.

The medium-range centroids are between $3.2\,\mathrm{km}$ and $6.8\,\mathrm{km}$ a.g.l. This variance is mostly proportional to the distance from CHC (Fig. 9a) with clusters 04_MR and 05_MR being slightly below the rest due to their location in the lowlands. Their height above sea level varies from 6.1 to $7.1\,\mathrm{km}$ (Fig. 9b). In terms of their S_{SRR}, these clusters vary from 7 to $20\,\%$. These values are approximately inversely proportional to their centroid distance from the station (Fig. 9c).

405   Finally, for the long-range subgroup, their clusters' centroid far distance from CHC is reflected in their mean height a.g.l. of $8.0\,\mathrm{km}$, mean height a.s.l. of $8.4\,\mathrm{km}$, and their low mean S_{SRR} of $0.7\,\%$. Furthermore, due to this high altitude and low influence from the surface, air masses arriving from these clusters are likely to present free troposphere characteristics. These clusters are all located west of the station (clock direction 07-10).

The temporal evolution of the clusters is shown in Figs. 10 and 11. All clusters within the short-range subgroup show a high

410   degree of temporal variability (Fig. 10a and b). Cluster 02_SR presents a high frequency variability which upon further analysis (via Fourier transform, Fig. S3) is shown to be a clear diurnal pattern. However, this pattern does not happen everyday (Fig. S4) but in 81 out of the 176 modelled days ($46\,\%$). During these days, the peak happens in the early afternoon (13h local time). The cluster 04_SR is highly variable as well however it does not present a diurnal pattern (Fig. S3) and its variation is similar to that of 05_MR, probably due to the fact that both clusters belong to the same pathway. Cluster 07_SR does present a sharp diurnal

415   pattern in 41 ($23\,\%$) out of 176 modelled days (Fig. S5) peaking at 11h local time. We conjecture that the peak time difference
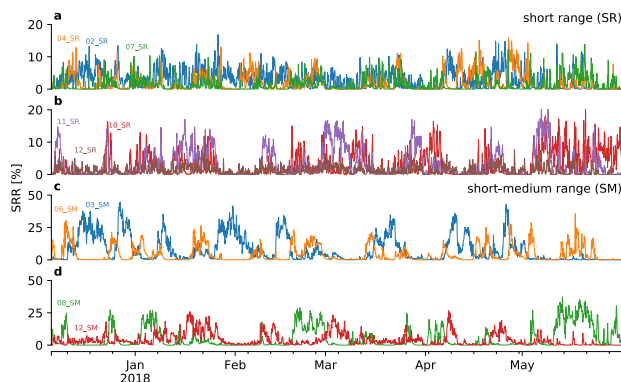
Figure 10. Time series of the SRR [%] cluster influence from 2017-12-06 until 2018-05-31. Panels a) and b) show the short-range (SR) clusters, c) and d) show the short-medium-range (SM) clusters. Note the different y-axis between panels a), b) and c), d).

between 02_SR (13h) and 07_SR (11h) is due to the different land type; 02_SR originates from the high humidity biome 1 (Tropical and subtropical forest, Fig. 5f) whereas 07_SR originates from the less humid biome 4 (Montane grasslands). This difference would entail different thermal inertia, different diurnal cycles and partitioning of the sensible and latent heat fluxes and thus a different boundary layer evolution. Cluster 07_SR is of particular interest since intense anthropogenic emission

420 sampled at CHC would most likely be generated in this highly populated area. Furthermore, the close contact of cluster 07_SR with the surface and its diurnal variability favour transport of emissions to the station during the day when PBL air from La Paz and El Alto is advected upslope by thermally driven winds. The cluster 10_SR, which originates close to Lake Titicaca, presents a diurnal pattern in 63 (36 %) out of 176 days peaking at 8h local time (Fig. S6). We conjecture that this early morning peak is related to the lake breeze circulation that develops due to the temperature difference between land and the lake. The

425 average surface temperature of the lake, obtained from Pillco Zolá et al. (2019), is $10\,°C$ and does not have a diurnal cycle. At 8h local time, the lake's surface temperature is higher than the surroundings favouring a land breeze (airflow from land to the lake) near the surface and ascent over the lake. The return flow, near the top of the BL, potentially advects air masses from the lake to CHC. The cluster 11_SR does not present a diurnal variation (Fig. S3) and its influence seems to be mostly driven by the medium-range cluster 11_MR (Fig. 11). Cluster 12_SR does present a diurnal pattern (Fig. S7) in 48 days out of the

430 176 modelled days (27 %) with a peak around 10h local time. This earlier peak, in comparison to 07_SR, is most likely due to its close location to CHC (the closest one, $14\,km$) so that early morning irradiation would create favourable conditions for up-slope winds.

The short-medium (SM), medium (MR) and long (LR) range clusters' temporal variability is shown in Fig. 10c and d, and Fig. 11a to d. The power spectra of the SRR intensity for these clusters is shown in Fig. S3 and is between 1 and 2 weeks

435 except for 08_SM and 12_SM which, in addition, also show a small diurnal variability (Fig. S3).

The prevalence of some short-medium range clusters changes during the 6-month period. Clusters 03_SM, 12_SM, 02_MR, 04_MR and 05_MR occur regularly from December (wet season) to April (transition season) but cease influencing in May
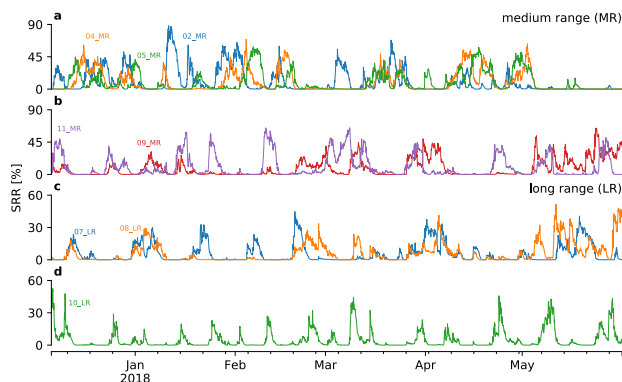
Figure 11. Time series of the SRR [%] cluster influence (similar to Fig. 10). Panels a) and b) show the medium range (MR) clusters. Panels c) and d) display the long range (LR) clusters.

(dry season). In contrast, the influence of clusters 06_SM, 11_MR and 10_LR do not change substantially during the 6 months whereas the influence of clusters 08_SM increases 09_MR, 07_LR and 08_LR in May.

440   Finally, in Fig. 12 we present a quantitative summary of the 18 clusters centroid properties along with their average SRR influence. Furthermore, we also link each of the 18 clusters to their main pathway. On average, there are 3 clusters per pathway and at least two distinct distance ranges. Furthermore, the clusters within each pathway are heterogeneous in position, surface influence and age which supports the idea that further insight into the air mass transport patterns can be attained by further subdividing the pathways.

445   5.4   An example application: sulfate from volcanic degassing

This section presents a proof of concept for our newly developed method to identify sources regions of air sampled at CHC. We use observations of particulate sulfate at CHC, performed with an Aerosol Chemical Speciation Monitor (ACSM) instrument, together with the results of our air mass history clustering analysis presented above to do so. During March and April, satellite imagery showed that while the Ubinas volcano was not degassing, the Sabancaya volcano, located $400\,\mathrm{km}$ WNW from CHC,

450   was and thus there was a clear, known, almost point source of particulate sulfate. As there are no other comparable strong point sources of particulate sulfate in the domain of interest, we assume that when high levels of particulate sulfate were measured at CHC, the air mass passed through the area near the volcano. Therefore, there should be a high correlation between the time series of sulfate and the SRR time series for clusters originating near the volcano.

To determine if this is the case, we calculated the 'Pearson' correlation coefficients for all available measurements of sulfate

455   from the ACSM to the SRR time series of each of the 18 clusters and all of the 6 pathways which are shown in Fig. 13a.

The cluster and pathway with the highest coefficients are 09_MR (0.40) and 08_PW (0.42) and both correlations have a $p_{\mathrm{value}} < 0.001$. The horizontal locations of this cluster and pathway are shown in Fig. 13b and their corresponding timeseries are shown in Fig. 13d. The timeseries for the sulfate measurements is shown in Fig. 13c. The correlation in combination with

short name $n_c = 18$	SRR [%] $n_c = 18$	distance from CHC [km]	height above ground [km]	height above sea level [km]	$\frac{SRR_{<1.5km}}{SRR_{tot}}$ [%]	age [h]	main pathway	SRR [%] $n_c = 6$
03_SM	8.5	134	2.5	4.1	36	34	03_PW	29.0
02_MR	10.8	671	6.8	7.1	7	60		
04_MR	9.6	687	5.2	6.2	14	58		
04_SR	1.9	36	1.2	5.1	70	18	05_PW	11.4
05_MR	9.6	506	3.3	6.5	24	50		
06_SM	4.3	169	2.4	6.1	34	35	07_PW	13.9
07_LR	4.7	972	8.5	8.6	0	64		
08_LR	4.9	962	8.1	8.1	0	61		
07_SR	1.2	25	0.8	4.9	87	12	08_PW	17.7
10_SR	2.8	43	0.9	5.0	82	14		
08_SM	5.0	188	2.5	5.7	38	37		
09_MR	8.8	355	3.2	6.1	29	45		
11_SR	3.3	61	1.4	5.1	60	27	11_PW	19.6
11_MR	11.4	390	4.2	6.1	19	54		
10_LR	4.9	917	7.4	8.6	2	68		
12_SR	1.2	14	0.5	4.9	98	9	12_PW	8.4
02_SR	2.9	36	1.2	4.3	77	15		
12_SM	4.3	119	2.6	4.2	42	41		

Figure 12. Properties of the eighteen clusters ($n_c = 18$). The short name's digits refer to the clockwise direction of the centre of mass of each cluster. The letters refer to the range: SR = short range, SM = short-medium range, MR = medium range, and LR = long range. The SRR [%] column describes the average contribution of each cluster. We also show the distance from CHC, height above ground and height above sea level of each cluster's centre of mass. Furthermore, $\frac{SRR_{<1.5\,km}}{SRR_{total}}$ [%] shows the ratio between the SRR below $1.5\,km$ and the SRR summed over the full vertical column (SRR_{total}). The age column shows the weighted median time (in hours) required by the air masses to arrive from the respective cluster to CHC. The last two columns describe the results of clustering the 18 clusters into 6 clusters (main pathways, $n_c = 6$). The digits also refer to the clockwise direction. The last column adds up the SRR [%] of the cluster belonging to each main pathway. The colours in the first column correspond to the 6 pathways and associated colours shown in Figs. 5a and b, 6, and 7.

either the simplified (6 pathways) and the specific (18 clusters) clustering scheme correctly assigns the source region to the
460   location of the degassing Sabancaya volcano.

The coefficients from the pathways are quite clear; the highest value of 0.42 is at least twice as high as the next candidate, 07_PW. This clearly distinguishes 08_PW as the best candidate for the source of measured sulfate. The 18 clusters also assign the highest coefficient value correctly to 09_MR. However, 10_SR, 11_SR and 08_LR all have similarly high correlation coefficients. Therefore, while the results from 18 clusters are better at pinpointing the source region of sulfate, other regions
465   also appear as plausible candidates so there is a risk of overfitting when too many regions are considered. This may be a result of this source regions being entangled in terms of their temporal influence over the station. For example, in this particular case, the high correlation of 10_SR, 11_SR and 08_LR could be attributed to the fact that air masses that travel through the region of the volcano also have a time residence in the regions where 10_SR, 11_SR and 08_LR are located. This example shows that the clustering scheme is successful in identifying regions in its simplified version (6 pathways) and also in its more detailed
470   version (18 clusters) albeit care must be taken when drawing conclusions so that we do not overfit to the source regions. Finally, it should be noted that only one example is presented here. Future work will include the extension to additional case studies, for example, comparing the clusters to measurements of black carbon.
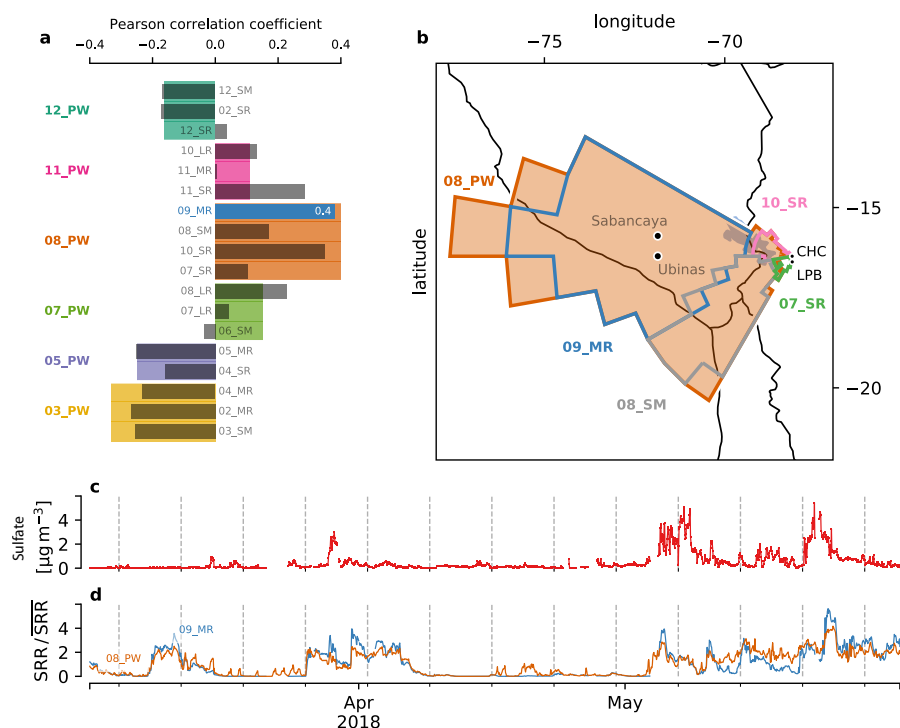
Figure 13. In panel a), we show the Pearson correlation coefficient between sulfate concentrations sampled at CHC and, both, the 18 clusters and the 6 pathways (PW). In panel b), the regions covered by the pathway 08_PW and the clusters 09_MR, 08_SM, 10_SR and 07_SR are shown along with Sabancaya and Ubinas volcanoes (both are located within 09_MR). All of the aforementioned clusters are contained by 08_PW. Panel c) and d show the timeseries of sulfate sampled at CHC and the normalized SRR timeseries of 08_PW and 09_MR respectively.

6 Discussion and recommendations

Previously, we described the characteristics and location of the 6 pathways and 18 clusters and related this to the surface type
475 (biomes). It is also relevant to consider how the pathway positions and time series relate to typical meteorological patterns.

During the austral summer, the Intertropical Convergence Zone (ITCZ) migrates south, coinciding with the decrease of the meridional gradient temperature and the associated southward shift of the westerly subtropical jet stream. At the same time, deep convection starts to develop especially in the central part of the continent. This favours the expansion of the equatorial easterly winds and thus a weak mean east-to-west flow in the middle and upper troposphere is established (Garreaud et al.,
480 2003). This east-to-west flow is well captured by 03_PW (03_SM, 02_MR and 04_MR) which is strong in DJFMA.

At the same time, the expansion of the trade winds also generates the South American Low Level Jet (SALLJ) which brings moisture from the Atlantic and the Amazon basin first to the eastern slopes of the Andes and then turning in a north-to-south fashion towards the southern part of the continent. It is known that the SALLJ is also responsible for bringing moisture to the Altiplano plateau since part of this flow is channelled into/goes over the plateau (Insel et al., 2010). This pattern correlates

485   well with 12_PW (12_SR, 02_SR and 12_SM) which brings low altitude air masses from the eastern slopes of the station. The pathway also is strong during DJFMA and diminishes its influence in May. It is important to note that 03_PW and 12_PW do not necessarily influence the station synchronously.

During the austral winter, the ITCZ migrates north and the subtropical westerly jet stream moves north reaching up to 20 degrees south Garreaud et al. (2003). This creates an upper-level, large-scale westerly flow that favours air masses from the
490   Pacific/Altiplano region. This coincides with the increment of the influence of both 07_PW (06_SM, 07_LR and 08_LR) and 08_PW (specially 08_SM and 09_MR). The first brings long-range subsiding air masses from high up in the troposphere (7.7 km a.s.l.) while the second advects dry air from Altiplano at low-level (2.5 km a.g.l.). Both of these clusters are present during the dry season and reach maximum intensity in May.

The connections of pathways 05 and 11 to the general atmospheric dynamics is not immediately evident. Pathway 05_PW,
495   which comes from the south, is strongest between December to March and has a low centroid height a.g.l. (3 km), could be linked to the development of a strong South Atlantic Convergence Zone starting in the eastern slopes of the Altiplano at latitudes of around 20° that would pull surface air from the south. On the other hand, 11_PW which consists of 3 clusters, originates from both from the Amazon and the Pacific (Fig. 8). These multiple sources areas, spanning both east and west, indicate that 11_PW is a "hybrid" pathway and therefore likely occurs in both the wet and dry seasons.

500   This study builds upon a previous source region analysis performed by Chauvigné et al. (2019) at the CHC station where a similar WRF setup was used in combination with back trajectories. Our methodological approach is notably different from this earlier study primarily as we use a Lagrangian dispersion transport model rather than a back trajectory model meaning that turbulent mixing and convection processes that air parcels experience during transport are better represented in this study. Additionally, our WRF simulations and hence meteorological data have a higher vertical resolution (61 levels compared to
505   28) and our 18 clusters provide more detail than the 6 clusters presented by Chauvigné et al. (2019). Our pathway results are largely in agreement with this previous study: similar source regions are observed for similar seasons and both studies show that source regions from the west start influencing the station in the transition month of May. However, key additions of this study are (1) the vertical distribution of the air masses sources is more accurately captured, (2) the influence of the surface and the pseudo PBL on air sampled at CHC is more accurately quantified and (3) the diurnal cycle is captured by the analysis.

510   The type of analysis performed here is applicable to many other stations worldwide both in mountainous regions but also for stations in non-mountainous areas which are equally influenced by local and remote sources. Therefore, based on the experience and knowledge gain here, we make the following recommendations for future studies:

– For source identification in regions with complex terrain we strongly recommend the use of Lagrangian dispersion models over simple, limited number back trajectories based approaches. This has been previously noted also by Stohl
515   et al. (2002).

– The accuracy of the meteorological input data is crucial for reliable results and therefore should be verified before performing the FLEXPART simulations. In our case, this step revealed large biases in the Titicaca lake temperature which affected local wind patterns.

– Selecting the optimal number of clusters is challenging and both quantitative scores and the scientific application should
520   be considered. We recommend that both pathways (smaller number) and clusters (larger number) are computed and their
centroid characteristics analysed. Both are useful and have notably different applications.

– Due to computational limitations, and given that the campaign only lasted for six months, we only simulated six months.
For a more complete physical understanding, if computational resources allow, we recommend that a full annual cycle
is always simulated even if the observational campaigns the model simulations support are shorter in duration.

525 **7   Summary and Conclusions**

In this study we successfully developed a new method to identify air mass source regions in sites of complex topography. We
then applied this methodology to the GAW station CHC, located near La Paz / El Alto at 5240 m.a.s.l. In order to accomplish
this, we started with a WRF simulation in combination with FLEXPART to create a high-resolution data set of source areas
for CHC. Then we applied our new method, based on cluster analysis, to transform the complex and large output dataset into
530 a user-friendly timeseries dataset of air mass source regions. We documented the characteristics of the identified source areas
and demonstrated the strength and simplicity of the method's classification results by applying our method to confirm that the
Sabancaya volcano is the source of sulfate measurements at the CHC station. The main conclusions of our analysis are:

– On average, 9% of the air sampled at CHC has been in contact with the surface, and 24% with the "pseudo" PBL, within
the previous 4 days. Therefore we can conclude that on average, at any given time $\sim$76% of the measured air mass at
535   CHC represents free tropospheric air. Thus, the air masses sampled at CHC are very rarely purely free tropospheric air
masses.

– The surface influence has a clear diurnal cycle, with low contributions during the night and higher contributions starting
at 10 am local time and continuing during the day. The duration of the high surface influence during daytime is longer
in the dry season (May) compared to the wet season (December - March).

540 – Air masses arriving at CHC have a wide range of sources covering many different biomes and altitudes and it is common
for any one specific sample time to have more than one source region.

– The most dominant pathway to emerge in our 6-month study is 03_PW which is responsible for 29% of the SRR and
originates in the Amazon. However, as we detected that this PW does not occur in May, we hypothesize, based on
Chauvigné et al. (2019), that if our analysis extended over all of the dry season (May-August), the overall prevalence of
545   this PW would decrease and others (e.g. 07_PW and 08_PW) would increase.

– For the clusters' centroid positions, a linear relationship exists between the horizontal distance from CHC and the height
above ground, with those farther away also being located higher in the atmosphere.

– Clusters located closest to CHC have the highest pseudo PBL influence and, rather than a linear decrease, the influence
of the pseudo PBL decreases almost exponentially with increasing distance from CHC.

550   – The contribution to the SRR is largest for the medium-range clusters and smallest for the short-range clusters thus showing no linear relationship with the distance from CHC.

    – The short-range clusters have high temporal frequency modulated by local meteorology driven by the diurnal cycle whereas the mid- and long-range clusters' variability occurs on timescales- governed by synoptic-scale dynamics.

To conclude, firstly, the method developed here can be applied to many other long term monitoring stations. Secondly, the
555 data sets produced here, that provide detailed information about the sources of air masses sampled at CHC, will be applied in forthcoming studies on the chemical composition measurements made at CHC during the SALTENA campaign.

Data availability. The SRR cluster timeseries are publicly available at https://doi.org/10.5281/zenodo.4539590. Both the WRF and FLEX-PART raw output datasets are available upon request.

Author contributions. DA performed the WRF and FLEXPART simulations and did the data analysis. EK corrected the lake temperature
560 problem and verified the WRF simulations. DA and VAS wrote the paper with contributions from all authors. SC and PA performed the sulfate measurements and the provided quality-controlled sulfate dataset used here. MA, PL, AW, RK contributed to the development of the GAW station CHC and to the overall planning for the SALTENA campaign. FB, VAS and RK conceived the study and led the overall scientific investigation.

Competing interests. The authors declare no competing interests.

565 *Disclaimer.* The authors declare no conflict of interest.

Acknowledgements.

    – We acknowledge CSC – IT Center for Science Ltd. for the generous allocation of computational resources.

    – DA and FB are funded by ERC (project CHAPAs no. 850614) and Finnish Centre of Excellence as well as the Academy of Finland (project no. 311932, 315203 and 337549).

570   – This study was supported by UMSA (Universidad Mayor de San Andrés) through the Laboratory for Atmospheric Physics (LFA) part of the Institute for Physics Research. The LFA provided scientific, administrative and logistical support.

    – We also acknowledge the financial, administrative and logistical support from IRD (Institut de Recherche pour le De´veloppement) representation in La Paz Bolivia, through funding from by Labex OSUG@2020.

    – PA acknowledges funding from FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo trough grant 2017/17047-0.

575 – DA wishes to thank the many scientists, technicians and personnel involved in the Chacaltaya station whose outstanding commitment enables high-quality atmospheric records in a challenging environment.

– DA thanks Sara Blichner for valuable input during the writing process.

References

Bianchi, F., Sinclair, V., Aliaga, D., Zha, Q., Scholz, W., Wu, C., Heikkinen, L., Modini, R., Partoll, E., Velarde, F., Moreno, I., Gramlich,
580   Y., Huang, W., Leiminger, M., Enroth, J., Perakyla, O., Marinoni, A., Chen, X., Blacutt, L., Forno, R., Gutierrez, R., Ginot, P., Uzu, G.,
      Facchini, M. C., Gilardoni, S., Gysek, M., Cai, R., Petaja, T., Rinaldi, M., Saathoff, H., Sellegri, K., Worsnop, D., Artaxo, P., Hansel, A.,
      Kulmala, M., Wiedensohler, A., Laj, P., Krejci, R., Carbone, S., Andrade, M., and Mohr, C.: The SALTENA Experiment: Comprehensive
      Observations of Aerosol Sources, Formation and Processes in the South American Andes, Bulletin of the American Meteorological
      Society, 2021, in review.

585 Bolaño-Ortiz, T. R., Pascual-Flores, R. M., López-Noreña, A. I., Ruggeri, M. F., Lakkis, S. G., Fernández, R. P., and Puliafito, S. E.:
      Assessment of Absorbing Aerosols on Austral Spring Snow Albedo Reduction by Several Basins in the Central Andes of Chile from Daily
      Satellite Observations (2000–2016) and a Case Study with the WRF-Chem Model, SN Appl. Sci., 1, 1352, https://doi.org/10.1007/s42452-
      019-1256-z, 2019.

Brattich, E., Hernández-Ceballos, M., Cinelli, G., and Tositti, L.: Analysis of 210Pb Peak Values at Mt. Cimone (1998–2011), Atmospheric
590   Environment, 112, 136–147, https://doi.org/10.1016/j.atmosenv.2015.04.020, 2015.

Brattich, E., Orza, J. A. G., Cristofanelli, P., Bonasoni, P., Marinoni, A., and Tositti, L.: Advection Pathways at the Mt. Cimone
      WMO-GAW Station: Seasonality, Trends, and Influence on Atmospheric Composition, Atmospheric Environment, 234, 117 513,
      https://doi.org/10.1016/j.atmosenv.2020.117513, 2020.

Brioude, J., Arnold, D., Stohl, A., Cassiani, M., Morton, D., Seibert, P., Angevine, W., Evan, S., Dingwell, A., Fast, J. D., Easter, R. C.,
595   Pisso, I., Burkhart, J., and Wotawa, G.: The Lagrangian Particle Dispersion Model FLEXPART-WRF Version 3.1, Geoscientific Model
      Development, 6, 1889–1904, https://doi.org/10.5194/gmd-6-1889-2013, 2013.

Brunner, D., Henne, S., Keller, C. A., Vollmer, M. K., Reimann, S., and Buchmann, B.: Estimating European Halocarbon Emissions Using
      Lagrangian Backward Transport Modeling and in Situ Measurements at the Jungfraujoch High-Alpine Site, in: LAGRANGIAN MOD-
      ELING OF THE ATMOSPHERE, edited by Lin, J., Brunner, D., Gerbig, C., Stohl, A., Luhar, A., and Webley, P., vol. 200 of *Geophysical*
600   *Monograph Book Series*, pp. 207+, https://doi.org/10.1029/2012GM001258, 2012.

Carneiro, R. G. and Fisch, G.: Observational Analysis of the Daily Cycle of the Planetary Boundary Layer in the Central Amazon
      during a Non-El Niño Year and El Niño Year (GoAmazon Project 2014/5), Atmospheric Chemistry and Physics, 20, 5547–5558,
      https://doi.org/10.5194/acp-20-5547-2020, 2020.

Cécé, R., Bernard, D., Brioude, J., and Zahibo, N.: Microscale Anthropogenic Pollution Modelling in a Small Tropical Island during Weak
605   Trade Winds: Lagrangian Particle Dispersion Simulations Using Real Nested LES Meteorological Fields, Atmospheric Environment, 139,
      98–112, https://doi.org/10.1016/j.atmosenv.2016.05.028, 2016.

Chauvigné, A., Sellegri, K., Hervo, M., Montoux, N., Freville, P., and Goloub, P.: Comparison of the Aerosol Optical Properties and Size
      Distribution Retrieved by Sun Photometer with in Situ Measurements at Midlatitude, Atmospheric Measurement Techniques, 9, 4569–
      4585, https://doi.org/10.5194/amt-9-4569-2016, 2016.

610 Chauvigné, A., Aliaga, D., Sellegri, K., Montoux, N., Krejci, R., Močnik, G., Moreno, I., Müller, T., Pandolfi, M., Velarde, F., Weinhold,
      K., Ginot, P., Wiedensohler, A., Andrade, M., and Laj, P.: Biomass Burning and Urban Emission Impacts in the Andes Cordillera Region
      Based on in Situ Measurements from the Chacaltaya Observatory, Bolivia (5240 m a.s.l.), Atmos. Chem. Phys., 19, 14 805–14 824,
      https://doi.org/10.5194/acp-19-14805-2019, 2019.

Cheng, F.-Y., Yang, Z.-M., Ou-Yang, C.-F., and Ngan, F.: A Numerical Study of the Dependence of Long-Range Transport of CO to a Mountain Station in Taiwan on Synoptic Weather Patterns during the Southeast Asia Biomass-Burning Season, Atmospheric Environment, 78, 277–290, https://doi.org/10.1016/j.atmosenv.2013.03.020, 2013.

Collaud Coen, M., Andrews, E., Aliaga, D., Andrade, M., Angelov, H., Bukowiecki, N., Ealo, M., Fialho, P., Flentje, H., Hallar, A. G., Hooda, R., Kalapov, I., Krejci, R., Lin, N.-H., Marinoni, A., Ming, J., Nguyen, N. A., Pandolfi, M., Pont, V., Ries, L., Rodríguez, S., Schauer, G., Sellegri, K., Sharma, S., Sun, J., Tunved, P., Velasquez, P., and Ruffieux, D.: Identification of Topographic Features Influencing Aerosol Observations at High Altitude Stations, Atmospheric Chemistry and Physics, 18, 12 289–12 313, https://doi.org/10.5194/acp-18-12289-2018, 2018.

Conen, F., Henne, S., Morris, C. E., and Alewell, C.: Atmospheric Ice Nucleators Active ≥ -12 °C Can Be Quantified on PM₁₀ Filters, Atmos. Meas. Tech., 5, 321–327, https://doi.org/10.5194/amt-5-321-2012, 2012.

Danielson, J. J. and Gesch, D. B.: Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010), USGS Numbered Series 2011-1073, U.S. Geological Survey, 2011.

de Foy, B., Fast, J. D., Paech, S. J., Phillips, D., Walters, J. T., Coulter, R. L., Martin, T. J., Pekour, M. S., Shaw, W. J., Kastendeuch, P. P., Marley, N. A., Retama, A., and Molina, L. T.: Basin-Scale Wind Transport during the MILAGRO Field Campaign and Comparison to Climatology Using Cluster Analysis, Atmospheric Chemistry and Physics, 8, 1209–1224, https://doi.org/10.5194/acp-8-1209-2008, 2008.

de Foy, B., Zavala, M., Bei, N., and Molina, L. T.: Evaluation of WRF Mesoscale Simulations and Particle Trajectory Analysis for the MILAGRO Field Campaign, ATMOSPHERIC CHEMISTRY AND PHYSICS, 9, 4419–4438, https://doi.org/10.5194/acp-9-4419-2009, 2009.

De Wekker, S. F. J. and Kossmann, M.: Convective Boundary Layer Heights over Mountainous Terrain a Review of Concepts, FRONTIERS IN EARTH SCIENCE, 3, https://doi.org/10.3389/feart.2015.00077, 2015.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E., Jones, B., Barber, C., Hayes, R., Kormos, C., Martin, V., Crist, E., and Saleem, M.: An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm, BioScience, 1, 1–12, https://doi.org/10.1093/biosci/bix014, 2017.

Fleming, Z. L., Monks, P. S., and Manning, A. J.: Review: Untangling the Influence of Air-Mass History in Interpreting Observed Atmospheric Composition, Atmospheric Research, 104-105, 1–39, https://doi.org/10.1016/j.atmosres.2011.09.009, 2012.

Fröhlich, R., Cubison, M. J., Slowik, J. G., Bukowiecki, N., Prévôt, A. S. H., Baltensperger, U., Schneider, J., Kimmel, J. R., Gonin, M., Rohner, U., Worsnop, D. R., and Jayne, J. T.: The ToF-ACSM: A Portable Aerosol Chemical Speciation Monitor with TOFMS Detection, Atmospheric Measurement Techniques, 6, 3225–3241, https://doi.org/10.5194/amt-6-3225-2013, 2013.

Garreaud, R., Vuille, M., and Clement, A. C.: The Climate of the Altiplano: Observed Current Conditions and Mechanisms of Past Changes, Palaeogeography, Palaeoclimatology, Palaeoecology, 194, 5–22, https://doi.org/10.1016/S0031-0182(03)00269-4, 2003.

Ghasemifard, H., Yuan, Y., Luepke, M., Schunk, C., Chen, J., Ries, L., Leuchner, M., and Menzel, A.: Atmospheric CO2 and Delta C-13 Measurements from 2012 to 2014 at the Environmental Research Station Schneefernerhaus, Germany: Technical Corrections, Temporal Variations and Trajectory Clustering, AEROSOL AND AIR QUALITY RESEARCH, 19, 657–670, https://doi.org/10.4209/aaqr.2018.01.0010, 2019.

Gratz, L. E., Jaffe, D. A., and Hee, J. R.: Causes of Increasing Ozone and Decreasing Carbon Monoxide in Springtime at the Mt. Bachelor Observatory from 2004 to 2013, Atmospheric Environment, 109, 323–330, https://doi.org/10.1016/j.atmosenv.2014.05.076, 2015.

650 Griffiths, A. D., Conen, F., Weingartner, E., Zimmermann, L., Chambers, S. D., Williams, A. G., and Steinbacher, M.: Surface-to-Mountaintop Transport Characterised by Radon Observations at the Jungfraujoch, Atmospheric Chemistry and Physics, 14, 12 763–12 779, https://doi.org/10.5194/acp-14-12763-2014, 2014.

Hanna, S. R.: Applications in Air Pollution Modeling, in: Atmospheric Turbulence and Air Pollution Modelling, edited by Nieuwstadt, F. T. M. and Dop, H., pp. 275–310, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-010-9112-1_7, 1984.

655 Herrmann, E., Weingartner, E., Henne, S., Vuilleumier, L., Bukowiecki, N., Steinbacher, M., Conen, F., Coen, M. C., Hammer, E., Jurányi, Z., Baltensperger, U., and Gysel, M.: Analysis of Long-Term Aerosol Size Distribution Data from Jungfraujoch with Emphasis on Free Tropospheric Conditions, Cloud Influence, and Air Mass Transport, Journal of Geophysical Research: Atmospheres, 120, 9459–9480, https://doi.org/10.1002/2015JD023660, 2015.

Hirdman, D., Burkhart, J. F., Sodemann, H., Eckhardt, S., Jefferson, A., Quinn, P. K., Sharma, S., Ström, J., and Stohl, A.: Long-Term Trends
660 of Black Carbon and Sulphate Aerosol in the Arctic: Changes in Atmospheric Transport and Source Region Emissions, Atmos. Chem. Phys., 10, 9351–9368, https://doi.org/10.5194/acp-10-9351-2010, 2010.

Insel, N., Poulsen, C. J., and Ehlers, T. A.: Influence of the Andes Mountains on South American Moisture Transport, Convection, and Precipitation, Clim Dyn, 35, 1477–1492, https://doi.org/10.1007/s00382-009-0637-1, 2010.

Kassomenos, P., Vardoulakis, S., Borge, R., Lumbreras, J., Papaloukas, C., and Karakitsios, S.: Comparison of Statistical Clustering Tech-
665 niques for the Classification of Modelled Atmospheric Trajectories, Theor Appl Climatol, 102, 1–12, https://doi.org/10.1007/s00704-009-0233-7, 2010.

Keresztesi, Á., Nita, I.-A., Birsan, M.-V., Bodor, Z., Pernyeszi, T., Micheu, M. M., and Szép, R.: Assessing the Variations in the Chemical Composition of Rainwater and Air Masses Using the Zonal and Meridional Index, Atmospheric Research, 237, 104 846, https://doi.org/10.1016/j.atmosres.2020.104846, 2020.

670 Laj, P., Klausen, J., Bilde, M., Plaß-Duelmer, C., Pappalardo, G., Clerbaux, C., Baltensperger, U., Hjorth, J., Simpson, D., Reimann, S., Coheur, P. F., Richter, A., De Mazière, M., Rudich, Y., McFiggans, G., Torseth, K., Wiedensohler, A., Morin, S., Schulz, M., Allan, J. D., Attié, J. L., Barnes, I., Birmili, W., Cammas, J. P., Dommen, J., Dorn, H. P., Fowler, D., Fuzzi, S., Glasius, M., Granier, C., Hermann, M., Isaksen, I. S. A., Kinne, S., Koren, I., Madonna, F., Maione, M., Massling, A., Moehler, O., Mona, L., Monks, P. S., Müller, D., Müller, T., Orphal, J., Peuch, V. H., Stratmann, F., Tanré, D., Tyndall, G., Abo Riziq, A., Van Roozendael, M., Villani, P.,
675 Wehner, B., Wex, H., and Zardini, A. A.: Measuring Atmospheric Composition Change, Atmospheric Environment, 43, 5351–5414, https://doi.org/10.1016/j.atmosenv.2009.08.020, 2009.

Laj, P., Bigi, A., Rose, C., Andrews, E., Lund Myhre, C., Collaud Coen, M., Lin, Y., Wiedensohler, A., Schulz, M., Ogren, J. A., Fiebig, M., Gliß, J., Mortier, A., Pandolfi, M., Petäja, T., Kim, S.-W., Aas, W., Putaud, J.-P., Mayol-Bracero, O., Keywood, M., Labrador, L., Aalto, P., Ahlberg, E., Alados Arboledas, L., Alastuey, A., Andrade, M., Artíñano, B., Ausmeel, S., Arsov, T., Asmi, E., Backman, J., Baltensperger,
680 U., Bastian, S., Bath, O., Beukes, J. P., Brem, B. T., Bukowiecki, N., Conil, S., Couret, C., Day, D., Dayantolis, W., Degorska, A., Eleftheriadis, K., Fetfatzis, P., Favez, O., Flentje, H., Gini, M. I., Gregorič, A., Gysel-Beer, M., Hallar, A. G., Hand, J., Hoffer, A., Hueglin, C., Hooda, R. K., Hyvärinen, A., Kalapov, I., Kalivitis, N., Kasper-Giebl, A., Kim, J. E., Kouvarakis, G., Kranjc, I., Krejci, R., Kulmala, M., Labuschagne, C., Lee, H.-J., Lihavainen, H., Lin, N.-H., Löschau, G., Luoma, K., Marinoni, A., Martins Dos Santos, S., Meinhardt, F., Merkel, M., Metzger, J.-M., Mihalopoulos, N., Nguyen, N. A., Ondracek, J., Pérez, N., Perrone, M. R., Petit, J.-E., Picard,
685 D., Pichon, J.-M., Pont, V., Prats, N., Prenni, A., Reisen, F., Romano, S., Sellegri, K., Sharma, S., Schauer, G., Sheridan, P., Sherman, J. P., Schütze, M., Schwerin, A., Sohmer, R., Sorribas, M., Steinbacher, M., Sun, J., Titos, G., Toczko, B., Tuch, T., Tulet, P., Tunved, P., Vakkari, V., Velarde, F., Velasquez, P., Villani, P., Vratolis, S., Wang, S.-H., Weinhold, K., Weller, R., Yela, M., Yus-Diez, J., Zdimal, V., Zieger,

P., and Zikova, N.: A Global Analysis of Climate-Relevant Aerosol Properties Retrieved from the Network of Global Atmosphere Watch (GAW) near-Surface Observatories, Atmospheric Measurement Techniques, 13, 4353–4392, https://doi.org/10.5194/amt-13-4353-2020, 2020.

Lloyd, S.: Least Squares Quantization in PCM, IEEE Transactions on Information Theory, 28, 129–137, https://doi.org/10.1109/TIT.1982.1056489, 1982.

Lopez, M., Schmidt, M., Ramonet, M., Bonne, J.-L., Colomb, A., Kazan, V., Laj, P., and Pichon, J.-M.: Three Years of Semicontinuous Greenhouse Gas Measurements at the Puy de Dôme Station (Central France), Atmospheric Measurement Techniques, pp. 3941–3958, https://doi.org/10.5194/amt-8-3941-2015, 2015.

Martin, S. T., Artaxo, P., Machado, L. a. T., Manzi, A. O., Souza, R. a. F., Schumacher, C., Wang, J., Andreae, M. O., Barbosa, H. M. J., Fan, J., Fisch, G., Goldstein, A. H., Guenther, A., Jimenez, J. L., Pöschl, U., Silva Dias, M. A., Smith, J. N., and Wendisch, M.: Introduction: Observations and Modeling of the Green Ocean Amazon (GoAmazon2014/5), Atmospheric Chemistry and Physics, 16, 4785–4797, https://doi.org/10.5194/acp-16-4785-2016, 2016.

Ng, N. L., Herndon, S. C., Trimborn, A., Canagaratna, M. R., Croteau, P. L., Onasch, T. B., Sueper, D., Worsnop, D. R., Zhang, Q., Sun, Y. L., and Jayne, J. T.: An Aerosol Chemical Speciation Monitor (ACSM) for Routine Monitoring of the Composition and Mass Concentrations of Ambient Aerosol, Aerosol Science and Technology, 45, 780–794, https://doi.org/10.1080/02786826.2011.560211, 2011.

Ou-Yang, C.-F., Lin, N.-H., Lin, C.-C., Wang, S.-H., Sheu, G.-R., Lee, C.-T., Schnell, R. C., Lang, P. M., Kawasato, T., and Wang, J.-L.: Characteristics of Atmospheric Carbon Monoxide at a High-Mountain Background Station in East Asia, Atmospheric Environment, 89, 613–622, https://doi.org/10.1016/j.atmosenv.2014.02.060, 2014.

Ou-Yang, C.-F., Chang, C.-C., Wang, J.-L., Shimada, K., Hatakeyama, S., Kato, S., Chiu, J.-Y., Sheu, G.-R., and Lin, N.-H.: Characteristics of Summertime Volatile Organic Compounds in the Lower Free Troposphere: Background Measurements at Mt. Fuji, Aerosol Air Qual. Res., 17, 3037–3051, https://doi.org/10.4209/aaqr.2017.04.0144, 2017.

Paris, J.-D., Stohl, A., Ciais, P., Nédélec, P., Belan, B. D., Arshinov, M. Y., and Ramonet, M.: Source-Receptor Relationships for Airborne Measurements of CO_2, CO and O_3 above Siberia: A Cluster-Based Approach, Atmos. Chem. Phys., 10, 1671–1687, https://doi.org/10.5194/acp-10-1671-2010, 2010.

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, E.: Scikit-Learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Pillco Zolá, R., Bengtsson, L., Berndtsson, R., Martí-Cardona, B., Satgé, F., Timouk, F., Bonnet, M.-P., Mollericon, L., Gamarra, C., and Pasapera, J.: Modelling Lake Titicaca's Daily and Monthly Evaporation, Hydrology and Earth System Sciences, 23, 657–668, https://doi.org/10.5194/hess-23-657-2019, 2019.

Putero, D., Cristofanelli, P., Laj, P., Marinoni, A., Villani, P., Broquet, A., Alborghetti, M., Bonafè, U., Calzolari, F., Duchi, R., Landi, T., Verza, G., Vuillermoz, E., and Bonasoni, P.: New Atmospheric Composition Observations in the Karakorum Region: Influence of Local Emissions and Large-Scale Circulation during a Summer Field Campaign, Atmospheric Environment, 97, 75–82, https://doi.org/10.1016/j.atmosenv.2014.07.063, 2014.

Qie, G., Wang, Y., Wu, C., Mao, H., Zhang, P., Li, T., Li, Y., Talbot, R., Hou, C., and Yue, T.: Distribution and Sources of Particulate Mercury and Other Trace Elements in PM2.5 and PM10 atop Mount Tai, China, Journal of Environmental Management, 215, 195–205, https://doi.org/10.1016/j.jenvman.2018.03.050, 2018.

725 Rousseeuw, P. J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, Journal of Computational and Applied Mathematics, 20, 53–65, https://doi.org/10.1016/0377-0427(87)90125-7, 1987.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E.: The NCEP Climate Forecast System Version 2, J. Climate, 27, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1, 2013.

730 Salisbury, G., Monks, P. S., Bauguitte, S., Bandy, B. J., and Penkett, S. A.: A Seasonal Comparison of the Ozone Photochemistry in Clean and Polluted Air Masses at Mace Head, Ireland, Journal of Atmospheric Chemistry, 41, 163–187, https://doi.org/10.1023/A:1014202229304, 2002.

Sarvaiya, J., Patnaik, S., and Kothari, K.: Image Registration Using Log Polar Transform and Phase Correlation to Recover Higher Scale, JPRR, 7, 90–105, https://doi.org/10.13176/11.355, 2012.

735 Serafin, S., Adler, B., Cuxart, J., De Wekker, S. F. J., Gohm, A., Grisogono, B., Kalthoff, N., Kirshbaum, D. J., Rotach, M. W., Schmidli, J., Stiperski, I., Vecenaj, Z., and Zardi, D.: Exchange Processes in the Atmospheric Boundary Layer over Mountainous Terrain, ATMO-SPHERE, 9, https://doi.org/10.3390/atmos9030102, 2018.

Spalding, M. D., Fox, H. E., Allen, G. R., Davidson, N., Ferdaña, Z. A., Finlayson, M., Halpern, B. S., Jorge, M. A., Lombana, A., Lourie, S. A., Martin, K. D., McManus, E., Molnar, J., Recchia, C. A., and Robertson, J.: Marine Ecoregions of the World: A Bioregionalization

740 of Coastal and Shelf Areas, BioScience, 57, 573–583, https://doi.org/10.1641/B570707, 2007.

Stohl, A., Eckhardt, S., Forster, C., James, P., Spichtinger, N., and Seibert, P.: A Replacement for Simple Back Trajectory Calculations in the Interpretation of Atmospheric Trace Substance Measurements, Atmospheric Environment, 36, 4635–4648, https://doi.org/10.1016/S1352-2310(02)00416-8, 2002.

Sturm, P., Tuzson, B., Henne, S., and Emmenegger, L.: Tracking Isotopic Signatures of CO2 at the High Altitude Site Jungfraujoch with Laser

745 Spectroscopy: Analytical Improvements and Representative Results, Atmos. Meas. Tech., 6, 1659–1671, https://doi.org/10.5194/amt-6-1659-2013, 2013.

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., and Woods, E.: Tslearn, A Machine Learning Toolkit for Time Series Data, Journal of Machine Learning Research, 21, 1–6, 2020.

Tositti, L., Riccio, A., Sandrini, S., Brattich, E., Baldacci, D., Parmeggiani, S., Cristofanelli, P., and Bonasoni, P.: Short-Term

750 Climatology of PM10 at a High Altitude Background Station in Southern Europe, Atmospheric Environment, 65, 142–152, https://doi.org/10.1016/j.atmosenv.2012.10.051, 2013.

Ubl, S., Scheringer, M., and Hungerbuhler, K.: Relationships between Atmospheric Transport Regimes and PCB Concentrations in the Air at Zeppelin, Spitsbergen, ENVIRONMENTAL SCIENCE & TECHNOLOGY, 51, 9784–9791, https://doi.org/10.1021/acs.est.7b02571, 2017.

755 von Engeln, A. and Teixeira, J.: A Planetary Boundary Layer Height Climatology Derived from ECMWF Reanalysis Data, Journal of Climate, 26, 6575–6590, https://doi.org/10.1175/JCLI-D-12-00385.1, 2013.

Wiedensohler, A., Andrade, M., Weinhold, K., Müller, T., Birmili, W., Velarde, F., Moreno, I., Forno, R., Sanchez, M., Laj, P., Ginot, P., Whiteman, D., Krejci, R., Sellegri, K., and Reichler, T.: Black Carbon Emission and Transport Mechanisms to the Free Troposphere at the La Paz/El Alto (Bolivia) Metropolitan Area Based on the Day of Census (2012), Atmospheric Environment, 194, 158–169,

760 https://doi.org/10.1016/j.atmosenv.2018.09.032, 2018.