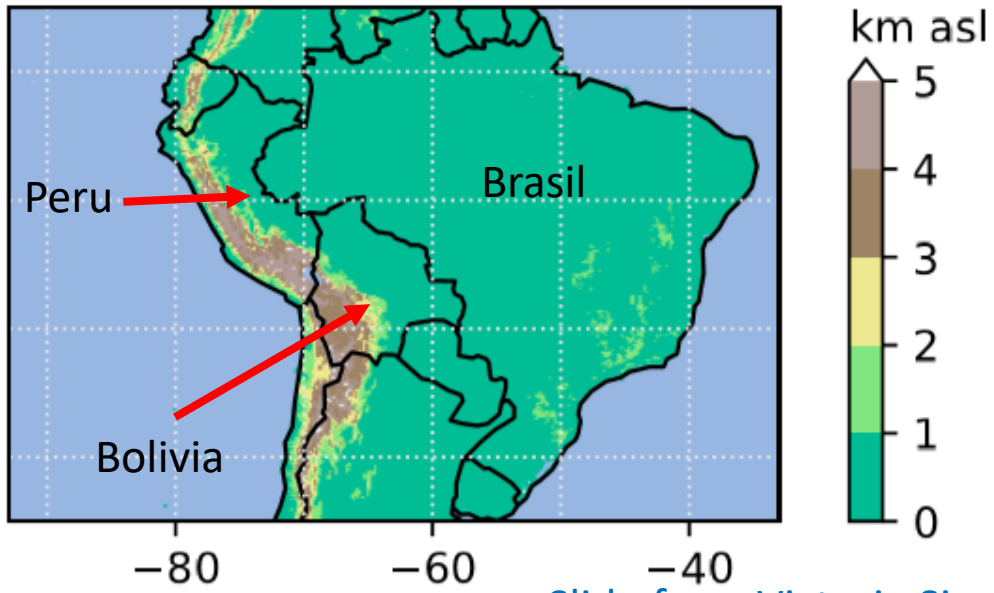
Source region cluster analysis at Chacaltaya with WRF and FLEXPART

- Diego Aliaga1, Victoria Sinclair1, Zha Qiaozhi1, Marcos Andrade2 Claudia Mohr3, Radovan Krejci3, and Federico Bianchi1
- 1Institute for Atmospheric and Earth System Research / Physics, University of Helsinki, Helsinki, Finland
- 2Laboratorio de Física de la Atmósfera, Universidad Mayor de San Andrés, La Paz, Bolivia
- 3Department of Environmental Science and Analytical Chemistry, Stock- holm University, Stockholm, Sweden

Bolivia Campaign

- **S**outhern hemisphere high **ALT**itude **E**xperiment on particle **N**ucleation **A**nd growth (SALTENA)
- An intensive campaign aiming to understand the nucleation characteristics at Chacaltaya
  - GAW station, 5240 m
- 6 months: December 2017 – May 2018
- State-of-the-art aerosol and trace gas instrumentation



Slide from Victoria Sinclair

Can we develop a method to identify the source regions of air at CHC?

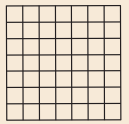
modeling

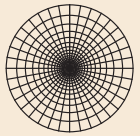
1. WRF — coarse reanalysis data

2. FLEXPART — high-res. meteo data

preprocessing

3. log-polar grid trans-formation

SRR_{xyzt} 

$\mathrm{SRR}_{r\theta zt}$ 

4. grid cell preprocessing

- gaussian filter (GF)

$$\mathrm{SRR}^{\mathrm{GF}}_{r\theta zt} = \mathrm{GF}(\mathrm{SRR}_{r\theta zt})$$

- quantile transform (QT)

$$\mathrm{SRR}^{\mathrm{QT}}_{r\theta zt} = \mathrm{QT}(\mathrm{SRR}^{\mathrm{GF}}_{r\theta zt})$$

- remove low influence cells

$$\mathrm{SRR}^{*}_{r\theta zt} = \left\{ \mathrm{srr} \in \mathrm{SRR}^{\mathrm{QT}}_{r\theta zt} \mid \mathrm{srr} > T \right\}$$

clustering

5. iterative k-means clustering algorithm (KCA)

$$K_{n_c} = \{C_i\}_{i=1}^{n_c} = \mathrm{KCA}\left(\mathrm{SRR}^{*}_{r\theta zt}, n_c\right) \quad \text{for } n_c \in \{2, 3, \ldots, 24\}$$

6. silhouette scoring (SS)

$$\mathrm{SS}(K_{n_c}) \quad \text{for } n_c \in \{2, 3, \ldots, 24\}$$

7. clustering results

$$\text{for } n_c = 18 \text{ and } n_c = 6 \quad K_{n_c} = \{C_i\}_{i=1}^{n_c}$$

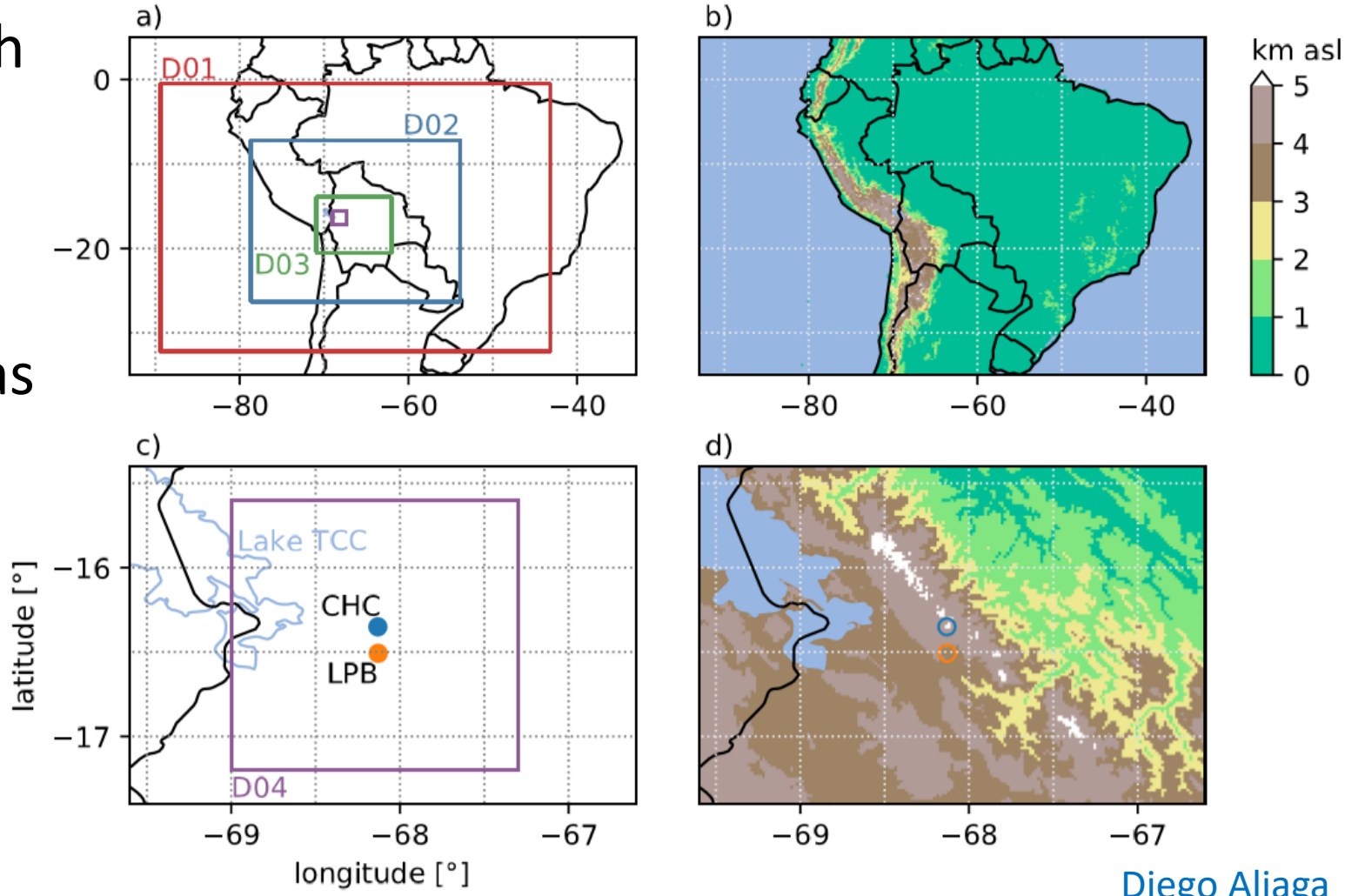
Applications

8. Vulcano — Pearson correlation method

2. ACSM data — Elsatic Net method

1 . WRF Simulations

- 6 month WRF simulation with nudging
- Output meteorological variables every 10 minutes
- 4 domains – inner domain has a resolution of 1 km.



Diego Aliaga

2 . FLEXPART Simulations

- Backward simulations every hour from Chacaltaya
- Release 20 000 particles every hour at a height of 10 m
- Particles advected backwards for 96 hours (4 days)

Diego Aliaga

2 . FLEXPART Simulations

- Backward simulations every hour from Chacaltaya
- Release 20 000 particles every hour at a height of 10 m
- Particles advected backwards for 96 hours (4 days)
- Obtain the source-receptor relationship (SRR, seconds)
  - Value for each grid cell (x,y,z) as a function of “arrival” time
  - SRR is related to the particles’ residence time in the output grid cells
  - Same SRR if one particle is present for 1 hour or if 2 particles are present for 0.5 hour (per unit mass)

Arrival time = time air mass was sampled at the station

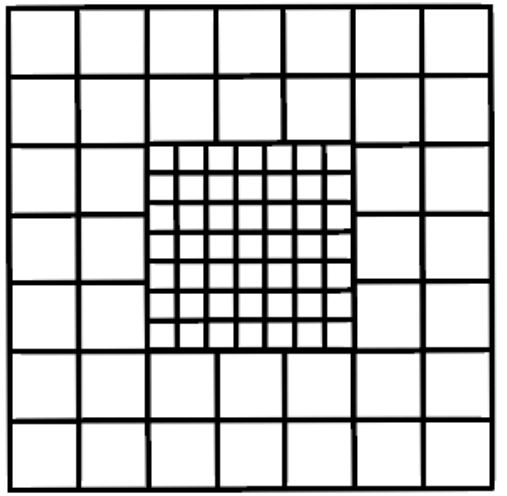
Receptor = Chacaltaya station in our case

Source= the physical source (not where the particles were released from!)

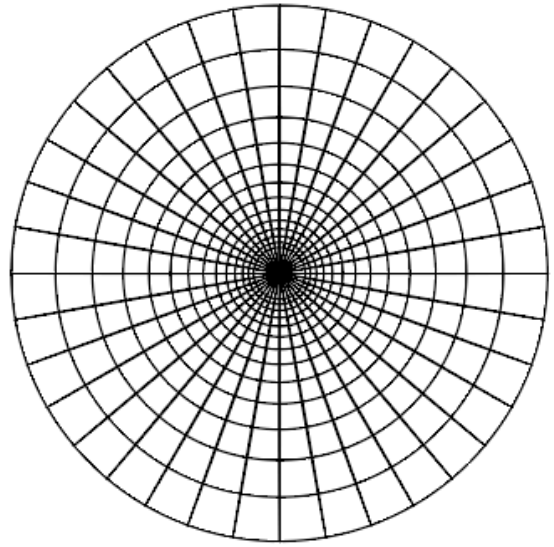
3. Log polar grid transformation

- Increased resolution over the sampling stations, decreased resolution for regions far away
- Gradual increment/decrement in the resolution (no kinks for the data analysis)
- Reduced dataset storage space:
  - **60 to 0.6 gigabytes**
  - Data analysis is also much faster

$$\mathrm{SRR}_{xyzt}$$



$$\mathrm{SRR}_{r\theta zt}$$



4. Grid cell preprocessing

- gaussian filter (GF)

$$\mathrm{SRR}^{\mathrm{GF}}_{r\theta zt} = \mathrm{GF}(\mathrm{SRR}_{r\theta zt})$$

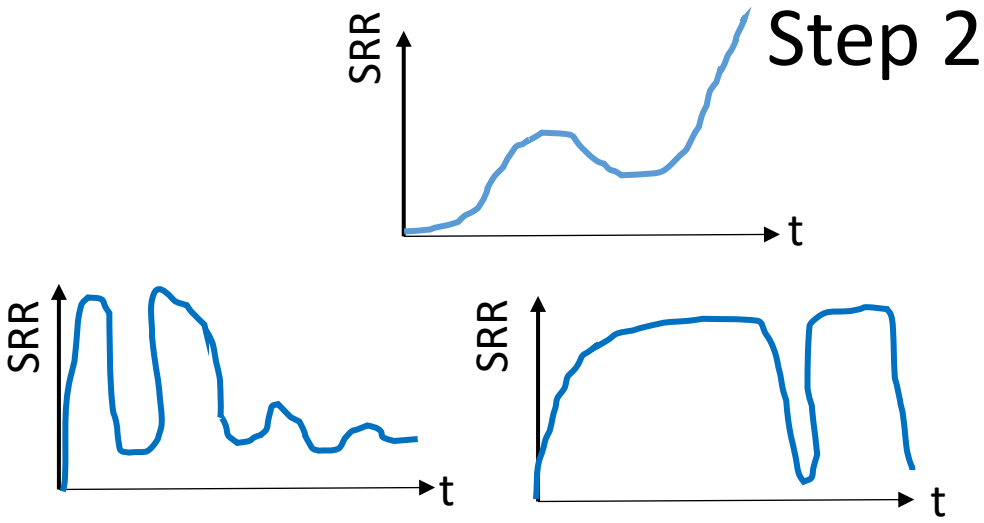
- quantile transform (QT)

$$\mathrm{SRR}^{\mathrm{QT}}_{r\theta zt} = \mathrm{QT}(\mathrm{SRR}^{\mathrm{GF}}_{r\theta zt})$$

- remove low influence cells

$$\mathrm{SRR}^{*}_{r\theta zt} = \left\{ \mathrm{srr} \in \mathrm{SRR}^{\mathrm{QT}}_{r\theta zt} \mid \mathrm{srr} > T \right\}$$

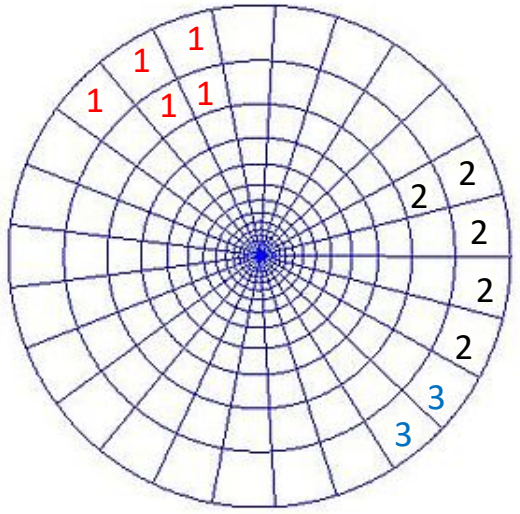
5. Clustering Method



Perform k means clustering on time series from each grid point -> output is a label of cluster number for each grid cell

For each grid cell belonging to each cluster sum SSR -> output cluster total SSR (seconds) as a function of time

For each “arrival” time get the total SSR over all clusters. For each cluster divide the cluster SSR by the total SSR -> output % contribution to total SSR of each cluster



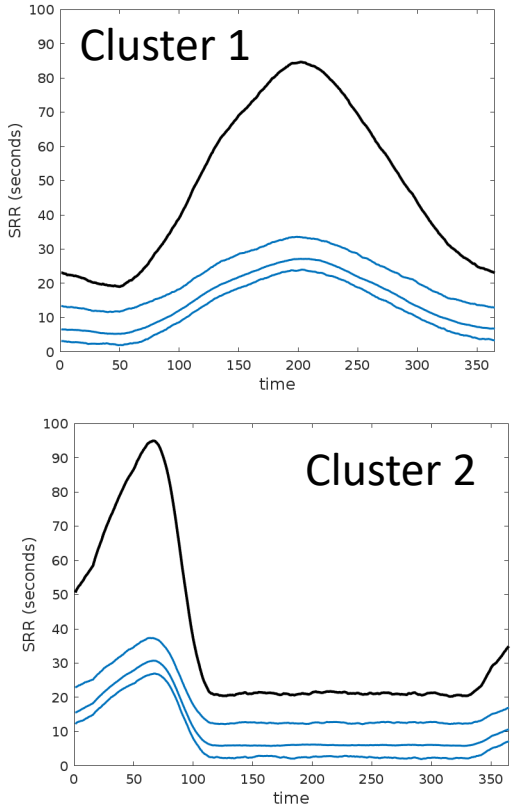
Number of clusters = 18. Decided based on silhouette score and to be relevant to our problem

5. Clustering Method

Perform k means clustering on time series from each grid point -> output is a label of cluster number for each grid cell

For each grid cell belonging to each cluster sum SSR -> output cluster total SSR (seconds) as a function of time

For each “arrival” time get the total SSR over all clusters. For each cluster divide the cluster SSR by the total SSR -> output % contribution to total SSR of each cluster



Two ”idealised” examples of SSR timeseries

Blue: individual grid points belonging to the cluster

Black: sum of all grid boxes in that cluster

Diego Aliaga

5. Clustering Method

SSR of cluster n at time t.

$$SRR_{pc}(n,t) = \frac{SRR_n(t)}{\sum_{n=1}^{n=18} SRR_n(t)} \times 100$$

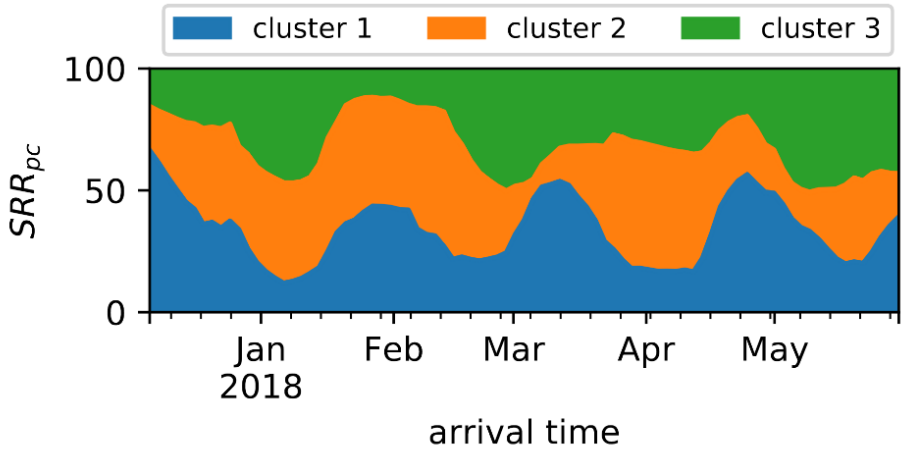
Sum of SSR over all 18 clusters at time t

Perform k means clustering on time series from each grid point -> output is a label of cluster number for each grid cell

For each grid cell belonging to each cluster sum SSR -> output cluster total SSR (seconds) as a function of time

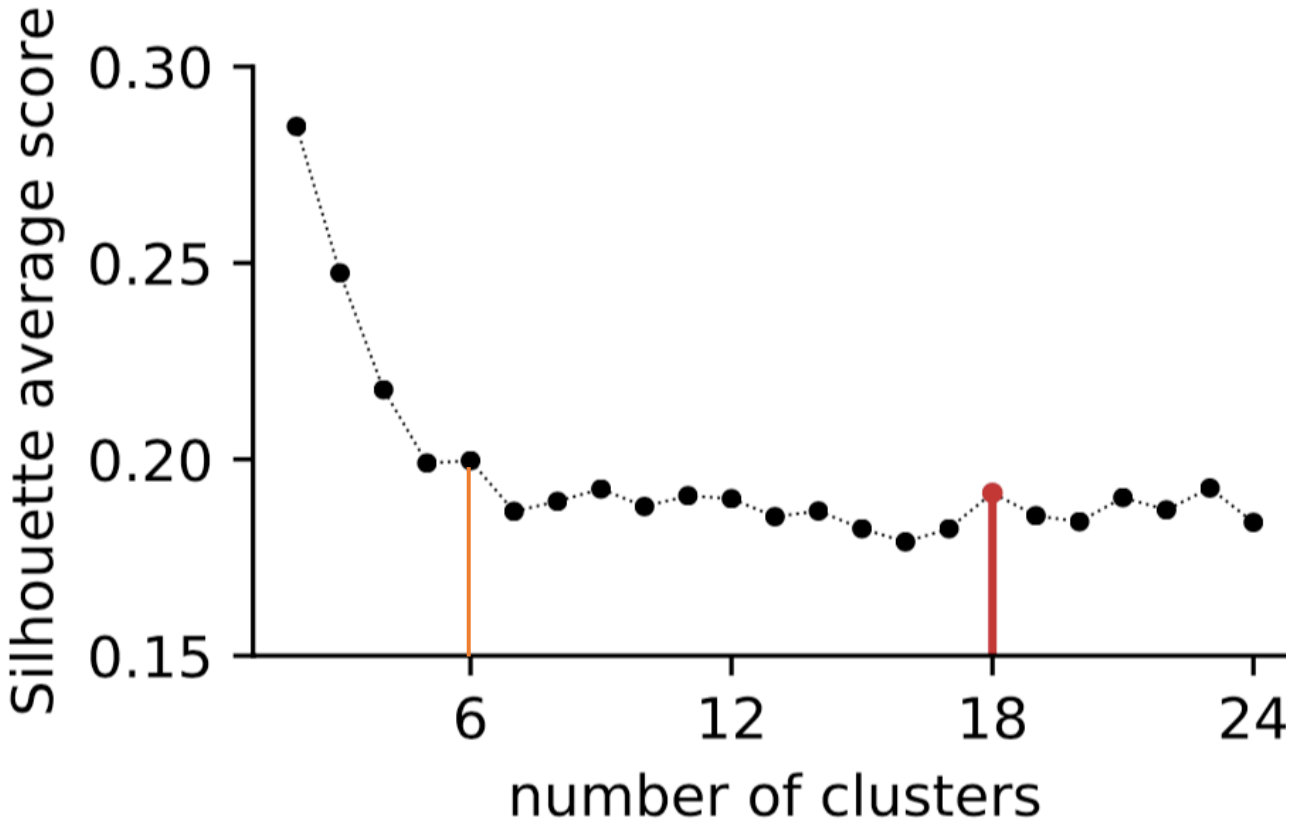
For each “arrival” time get the total SSR over all clusters. For each cluster divide the cluster SSR by the total SSR -> output % contribution to total SSR of each cluster

Example of SRR_{pc} for $n = 3$ clusters

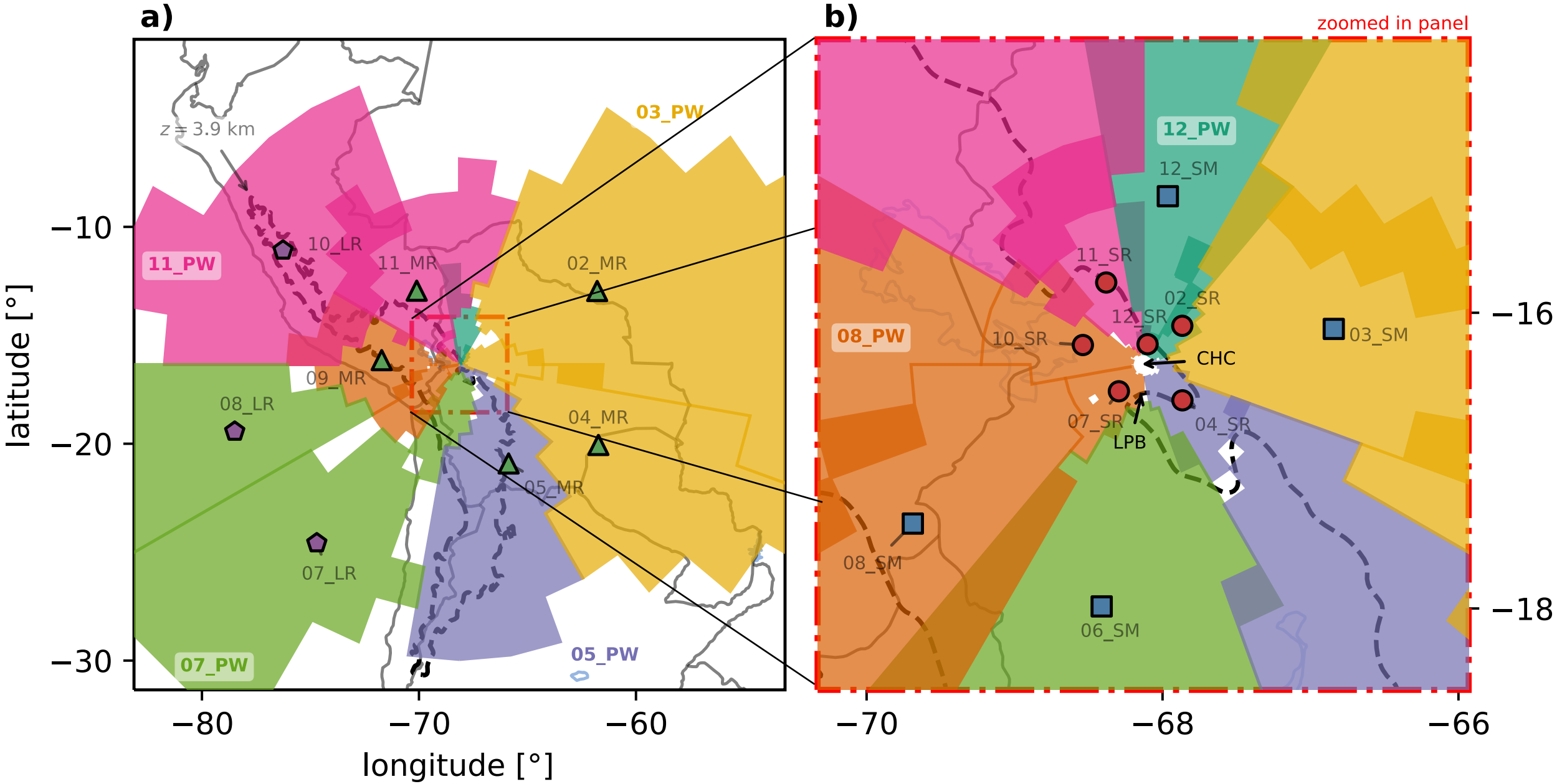


Diego Aliaga

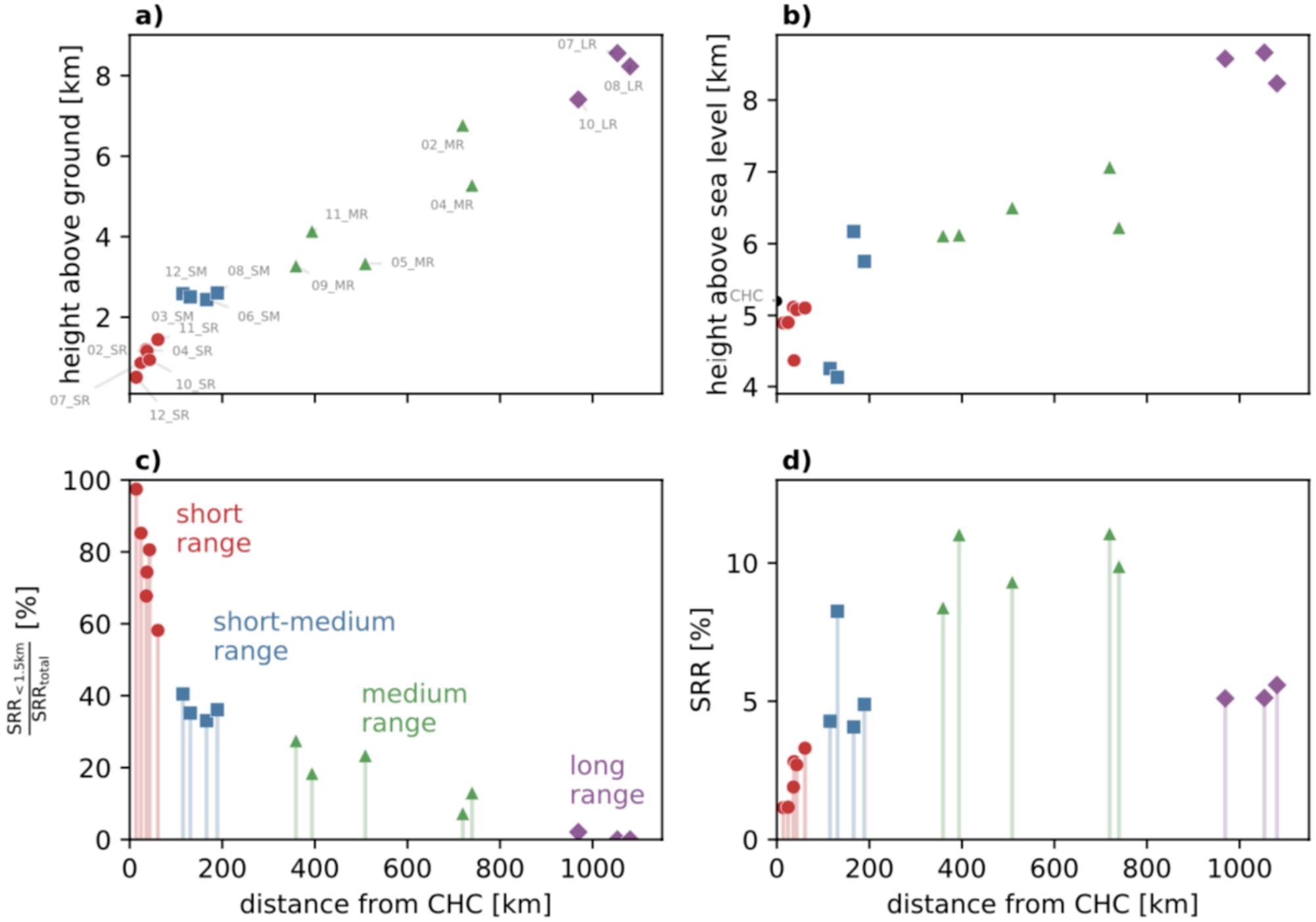
6. Silhouette Scoring



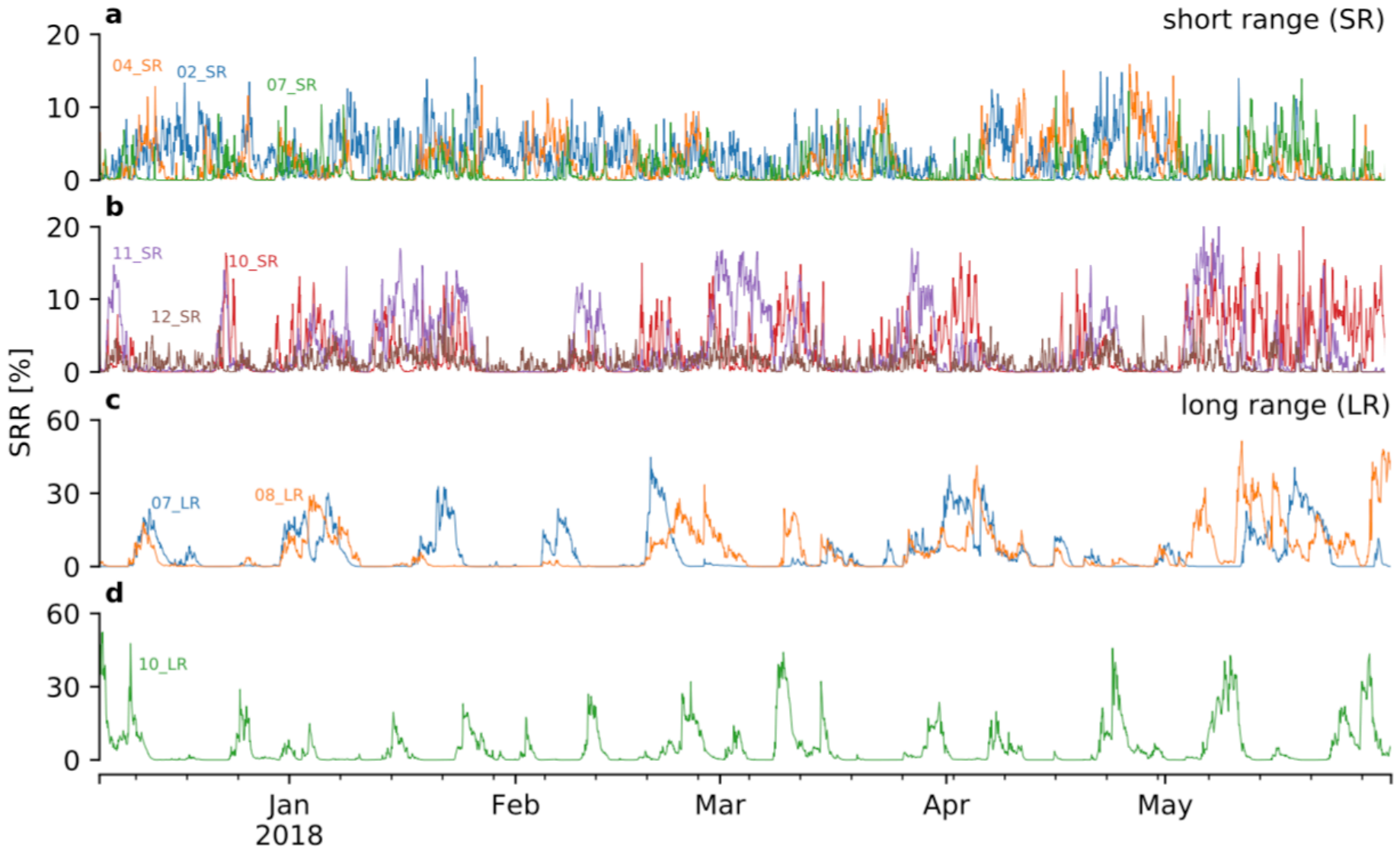
7. Clustering results



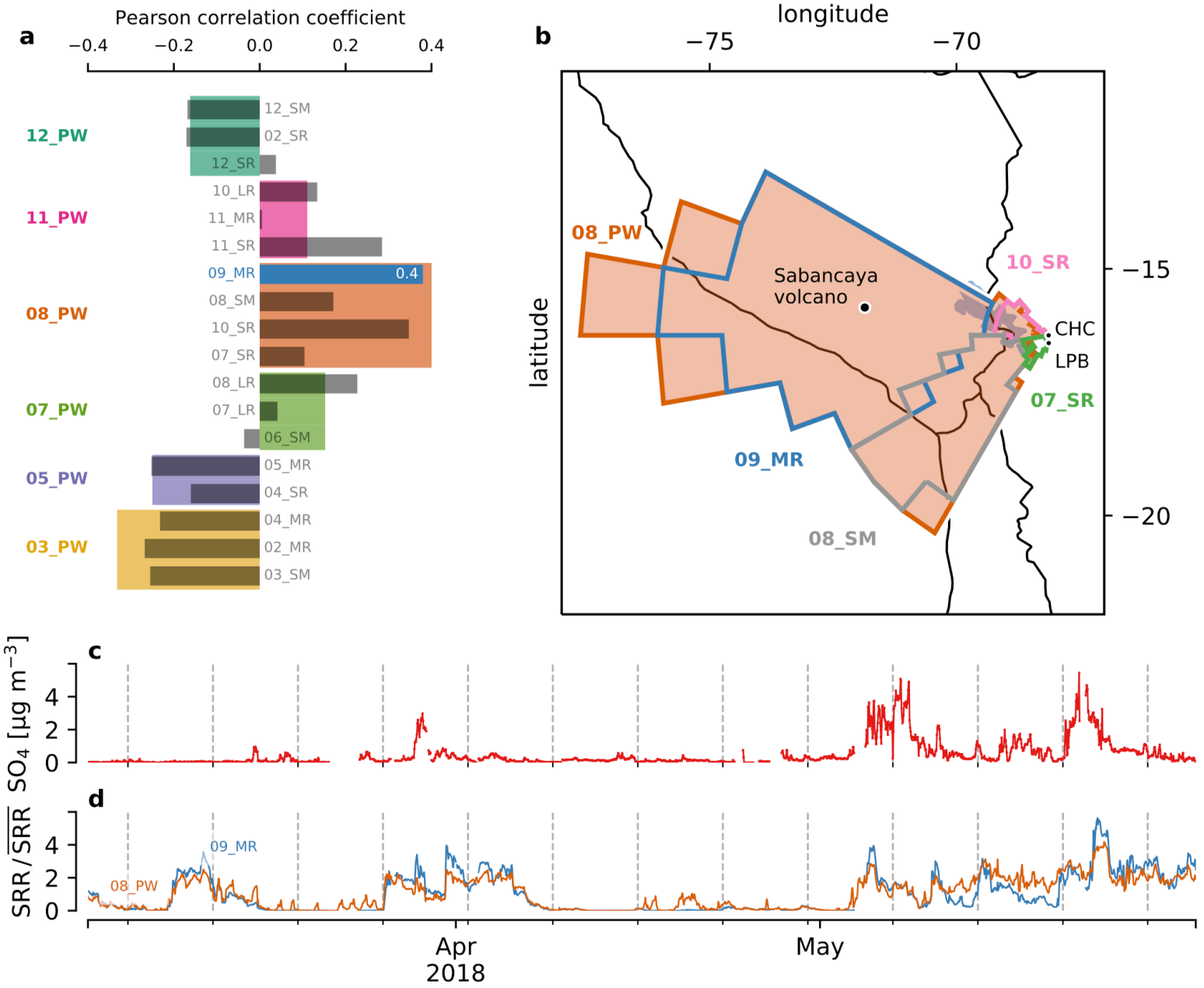
7. Clustering results



7. Clustering results



8. Vulcano correlation between sampled SO4 at Chacaltaya station and the SRR clusters



Conclusions

- A clustering method using SRR timeseries can identify source areas of air masses sampled at Chacaltaya
- We can quantify the various air mass influences at Chacaltaya
- We can identify source regions for SO4 (Volcano), Nitrates, Organics (La Paz)
- Future plans: use cluster data to understand other observations