Low Dimensional Embedding of Climate Data for Radio Astronomical Site Testing in the Colombian Andes

Germán Chaparro-Molano[1★], Oscar Leonardo Ramírez-Suárez[1], Oscar Restrepo[1,2], and Alexander Martínez[1,3]

[1] *Grupo de Simulación, Análisis y Modelado, Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia*
[2] *Radio Astronomy Instrumentation Group, Universidad de Chile, Santiago de Chile, Chile*
[3] *Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá, Colombia*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We set out to evaluate the potential of the Colombian Andes for millimeter-wave astronomical observations. Previous studies for astronomical site testing in this region have suggested that nighttime humidity and cloud cover conditions make most sites unsuitable for professional visible-light observations. Millimeter observations can be done during the day, but require that the precipitable water vapor column above a site stays below ∼10 mm. Due to a lack of direct radiometric or radiosonde measurements, we present a method for correlating climate data from weather stations to sites with a low precipitable water vapor column. We use unsupervised learning techniques to low-dimensionally embed climate data (precipitation, rain days, relative humidity, and sunshine duration) in order to group together stations with similar long-term climate behavior. The data were taken over a period of 30 years by 2046 weather stations across the Colombian territory. We find 6 regions with unusually dry, clear-sky conditions, ranging in elevations from 2200 to 3800 masl. We evaluate the suitability of each region using a quality index derived from a Bayesian probabilistic analysis of the station type and elevation distributions. Two of these regions show a high probability of having an exceptionally low precipitable water vapor column. We compared our results with global precipitable water vapor maps and find a plausible geographical correlation with regions with low water vapor columns (∼ 15 mm) at an accuracy of ∼ 20 km. Our methods can be applied to similar datasets taken in other countries as a first step toward astronomical site evaluation.

Key words: atmospheric effects – methods: data analysis – methods: statistical – site testing

1 INTRODUCTION

The development of astronomical instrumentation technology in the 0.2-2 THz range has been rapidly growing in recent years. Another recent surge in interest in telecommunications THz technology has appeared recently due to the saturation of the "classical radio window" in telecommunications (Hosako et al. 2007). For this reason, atmospheric models for absorption of THz photons (Rosenkranz 1998; Slocum et al. 2013) and artifact removal models (Withayachumnankul et al. 2008) have been developed. Countries near the Equator face a challenge in using this frequency band for astronomical observations and telecommunications using terabit satellite links (Suen 2016) due to the presence of a tropical belt of dense water vapor which efficiently absorbs THz radiation (Seidel et al. 2008). In northern South America, previous studies of astronomical site testing in the visible range have shown that high nighttime humidity conditions make this region suitable only for educational observatories (Pinzón et al. 2015). However, considering that millimeter/sub-millimeter observations need not be done during the night, unusually dry, clear-daytime-sky, high elevation regions in the northern Andes could be suitable candidates for a millimeter-wave observatory.

Astronomical observations in the millimeter and sub-millimeter wavelength range require that atmospheric effects affecting absorption at these wavelengths are kept

[★] E-mail: gchaparrom@ecci.edu.co

to a minimum (Chamberlin & Grossman 2012; Radford & Peterson 2016; Cortés et al. 2016). The main factor contributing to the atmospheric opacity is water vapor, which very efficiently absorbs light in the THz range (Pardo et al. 2001a,b; Kuhn et al. 2002) due to a continuum absorption spectrum formed by collisionally broadened absorption lines of water vapor in this frequency range (Clough et al. 1989; Pickett et al. 1998; Turner et al. 2009). In order to characterize a site according to its atmospheric transparency to THz radiation, it is necessary to retrieve the precipitable water vapor profile either via remote sensing techniques such as microwave radiometry (Peter & Kämpfer 1992; Paine et al. 2000; Battistelli et al. 2012), satellite measurements (Aumann et al. 2003; Jones et al. 2012; Suen et al. 2014; Wentz & Meissner 2016), radiosonde humidity measurements (Liljegren et al. 2001; Luini & Riva 2016) or indirectly via models of in situ climatological measurements (Lew & Uscka-Kowalkowska 2016) or GPS-delay studies (Niell et al. 2001; Wang et al. 2007).

Most of the precipitable water vapor that can potentially affect the path of a THz photon hoping to get through the atmosphere actually exists in the troposphere. The lower the elevation of a potential site, the longer the path becomes for that photon, increasing its chances of being absorbed by a water vapor molecule (Liebe 1989; Clough et al. 1989; Slocum et al. 2013). For this reason, the quality of a site is determined by the atmospheric water vapor column above the potential site (Smith et al. 2001; Cimini et al. 2007; He et al. 2012; Bustos et al. 2014). This water vapor column is usually expressed as the amount of precipitable water vapor, given in mm.

The water vapor column above a potential site can change rapidly (during the course of the day) or it can change seasonally (Cadeddu et al. 2013; Caumont et al. 2016). This is why medium-to-long term monitoring of the atmosphere above a site is required before building an expensive mm/sub-mm wave radio telescope. However, even in best-case scenarios, earth-based telescopes will always be limited by atmospheric absorption of light specially at THz frequencies, where frequency windows of observations are few and in some cases, narrow (Archibald et al. 2002; Denny et al. 2013).

Although radiometer measurements are desirable for directly measuring the atmospheric water vapor column, their development and/or deployment can be complex and expensive (Pazmany 2007; Peng et al. 2009). For this reason we screened historic climate data for evidence of local long-term clear sky, low humidity conditions as proxies for a locally low precipitable water vapor column. By low-dimensionally embedding the data, we are able to reduce the dimensionality of multi-annual monthly data from $N_{\rm dim} = 12$ to $N_{\rm dim} \leq 3$ while preserving $> 95\%$ of the variance in the data. Thus, we can correlate unusually dry regions regardless of their geographical location. Given that the climatological variables were sparsely measured by weather stations across Colombia, we evaluated the aptness of a given location using a Bayesian probability-derived quality index. Finally, we geographically clustered our candidate locations in order to identify regions of interest. We compared these regions of interest with regions with unusually low precipitable water vapor in low-resolution ($\sim$ 20 km) water vapor satellite maps (Suen 2016) and find a possibly significative geographical correlation. In an upcoming paper we will analize satellite data in order to retrieve high-resolution seasonal precipitable water vapor maps for the regions of interest reported here.

This paper is organized as follows. We describe our dataset and the distribution and types of weather stations in Section 2. In Section 3 we describe how we applied low-dimensional embedding algorithms to our data. This is followed by a discussion in Section 4, where we include an overview of climate patterns in Colombia, and describe our Bayesian probabilistic quality index to assess the suitability of a given weather station. We summarize our results in Section 5, where we identify candidate regions of interest for a mm-wave astronomical observatory site. Finally, we discuss our main conclusions and future perspectives in Section 6.

2 THE DATASET

The meteorological data used in this study were gathered over 30 years (1981-2010) in 2046 weather stations monitored by the Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) in Colombia (Figure 1). The variables that we considered relevant to our work were: Elevation (in meters above sea level, or masl), Precipitation (mm/mo), Rain Days (d/mo), Relative Humidity (%), and Sunshine Duration (h/d). The climatological variables are reported as multi-annual monthly data, i.e. each variable is reported monthly from January to December averaged over the 30 year range. For precipitation, each monthly datum corresponds to the cumulative monthly value. For rain days, each monthly datum corresponds to the number of rain days for that month. For relative humidity, each monthly datum corresponds to the average daily value averaged over each month. For sunshine duration, each monthly datum corresponds to the daily value averaged over each month. All stations reported precipitation values, although only 2002 reported rain days, 445 reported relative humidity, and 336 reported sunshine duration. This means that not all stations registered all the meteorological variables relevant to this work. Thus, we classify stations according to which variables they measured (Table 1). The data is public and can be found in the repository for this paper at `https://github.com/saint-germain/ideam` .

3 LOW-DIMENSIONAL EMBEDDING OF CLIMATOLOGICAL DATA

Our goal is to classify and group together stations that show a similar climatological behavior for each variable, regardless of the station's location, elevation or type. Since climate data track multi-annual seasonal variations, we can correlate weather patterns of regions which are not obviously related climatologically but might show a similar behavior, e.g. dry weather patterns in Guajira (a desert region in the north of Colombia) with similar dry conditions

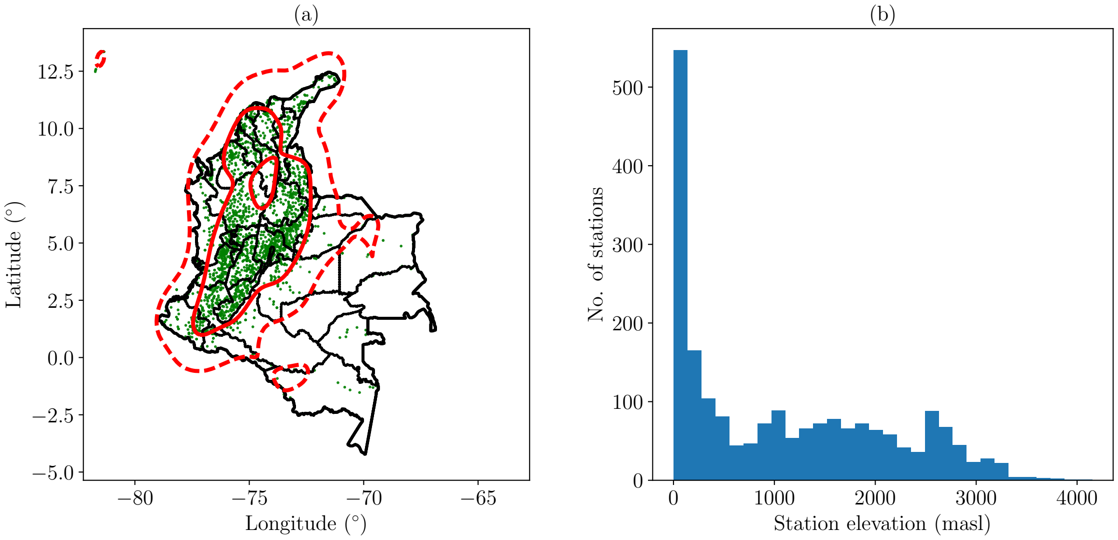

Figure 1. Distribution of weather stations in the dataset. (a) 1σ (solid) and 2σ (dashed) countours of a gaussian kernel density estimation of the geographical distribution of weather stations (small green points) from the IDEAM 1981-2010 database. The estimation was done using the Python `scipy.stats.gaussian_kde` function (Jones et al. 01). (b) Elevation histogram for all stations in the dataset.

Table 1. Types of stations classified according to its reported variables. Y = measured, N = not measured.

Type	Precip.	Rain Days	Relative Humidity	Sunshine Duration	No. of Stations
T_1	Y	N	N	N	29
T_2	Y	Y	N	N	1563
T_3	Y	N	Y	N	1
T_4	Y	Y	Y	N	117
T_5	Y	N	N	Y	1
T_6	Y	Y	N	Y	8
T_7	Y	N	Y	Y	13
T_8	Y	Y	Y	Y	314

present at high-mountain sites in the Andes.

We classified stations using a low-dimensional embedding (Shaw & Jebara 2009) of the available climate data. This ensures that a classification algorithm can uncover features in the data even when the data are projected across a number of dimensions that is lower than its native dimensionality (which in this case is monthly, i.e. $N_{\rm dim} = 12$). This involves performing a dimensionality reduction algorithm followed by a clustering algorithm. Given that we wish to make a grouping of stations showing similar climatological patterns without making assumptions on the data, we decided to use unsupervised learning techniques. In our case, we chose Principal Component Analysis (PCA) covering 95% of the variance followed by a Gaussian Mixture Model selected by a low/locally minimum Bayesian Information Criterion (BIC).

3.1 Principal Component Analysis

We were able to reduce the dimensionality of the data for each variable from $N_{\rm dim} = 12$ to $N_{\rm dim}^{\rm PCA} \leq 3$, while ensuring that at least 95% of the variance of the data is explained by the least amount of components. For precipitation, rain days and sunshine duration, $N_{\rm dim}^{\rm PCA}(95\%) = 3$ and for relative humidity $N_{\rm dim}^{\rm PCA}(95\%) = 2$. To do this we used the Python `sklearn.decomposition.PCA` module (Pedregosa et al. 2011). To illustrate, Figure 2 shows the relative humidity data projected across 2 principal components explaining 95% of the variance, thus preserving most of the original structure of the data. Dimensionality-reduced precipitation, rain days, and sunshine duration data are shown later (Figure 4).

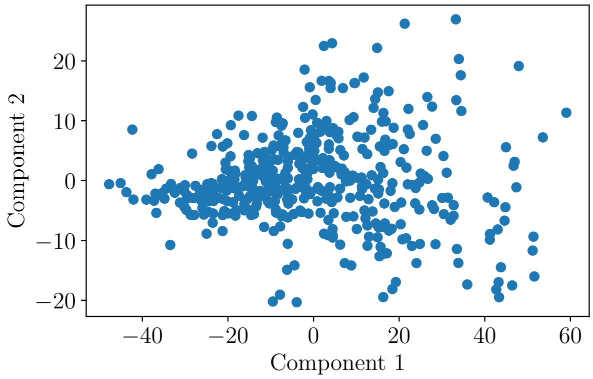

Figure 2. Relative humidity data projected across 2 principal components. The dimensionality of the data has been reduced from 12 to 2 while covering 95% of the variance of the data.

Table 2. Climate clusters according to a lowest BIC-based selection of Gaussian Mixture Models following Figure 3. The selected clusters column refers to the climate clusters shown in Figures 5 to 9.

Variable	Symbol	No. of clusters	Covariance Method	Selected Clusters
Relative Humidity	H	2	Full	2
Precipitation	R	11	Full	2,3,7,11
Rain Days	D	6	Full	1,4,6
Sunshine Duration	S	5	Tied	1,4,5

3.2 Gaussian Mixture Models

After projecting the data along the components found by the PCA algorithm, we clustered the dimensionally-reduced data using a Gaussian Mixture Model, which uses an Expectation-Maximization algorithm to find a Maximum Likelihood model composed of a number $N_{\mathrm{G-C}}$ of Gaussian distributions, each of which represents a cluster in the data (Reynolds 2009). Depending on the restrictions put on the covariance of each Gaussian distribution across the data space, the number of free parameters can go from $N_{\mathrm{G-C}}(N_{\mathrm{dim}}+1)$ (spherical covariance) to $N_{\mathrm{G-C}}(3N_{\mathrm{dim}}+1)$ (full covariance). Since the number of components and covariance restriction is given by the user and not *a posteriori* by the algorithm, we need to make sure that the Gaussian components do not over-fit the data, e.g. a model which yields one cluster per datum should be disallowed.

In order to achieve this, we compare the Bayesian Information Criterion (Schwarz 1978) for a grid of Gaussian clusters ($N_{\mathrm{G-C}} < 20$) and covariance restriction methods for each variable (Figure 3). To do this we used the Python `sklearn.mixture.GMM` module (Pedregosa et al. 2011). We selected the models which produced the lowest BIC value, and plotted the GMM-classified, PCA-projected climate data in Figure 4.

From the climatological clusters in Figures 5-12 we selected clusters that indicate a dry, sunny climate. Table 2 shows the lowest-BIC number of clusters and covariance method for each variable along with our selected clusters. From here on, if a station appears in one of our preferred clusters, we will say that it satisfies our criterion for that specific variable.

4 DISCUSSION

We wish to identify all stations for whom all four criteria (humidity, precipitation, rain days, sunshine duration) for identifying a region as having a stable sunny, dry climate can be met. However, since only a few stations measured all variables, it is not straightforward to accept or reject stations based on their satisfying a criterion, i.e. belonging to one of our preferred clusters. This requires setting up a measure of quality for individual stations depending on their location and the variables they measured. We will describe the selection process and probabilistic analysis that allows us to quantify a station's capability for indicating regions with an unusually dry, sunny climate.

We start by creating a shortlist of stations that meet criteria for (and only for) the variables they measured. This means that if a station has precipitation and humidity data, it should belong to a humidity and precipitation cluster listed in Table 2 or else it is rejected. 665 stations across Colombia showed such behavior. However, since we are interested in high-mountain regions, we included a criterion for elevation limiting our list to 119 stations located above 2000 masl. This shortlist was further reduced by requiring that the total average plus 1σ for precipitation, rain days, and relative humidity is less than 4.3 mm/d, 17 d/mo, and 81% respectively, and for sunshine duration, more than 4 h/d for any given station. This left us with 83 weather stations.

However, we want to give a measure of the quality of the station vis-a-vis our criteria, i.e. a station that meets less than all four criteria should still be considered in our analysis, but it cannot be given the same importance as a station that meets all four criteria. In order to account for this, and controlling for the fact that stations are not distributed evenly in elevation (Figure 1b), we propose a quality measure based on a probabilistic analysis of the data.

4.1 Bayesian probabilistic analysis

We quantify the quality of a given site using the probability for the hypothesis that a station j which appears in our shortlist meets all four criteria, given that it is located at a given elevation. We can write down this probability P_j in terms of the following events:

- C is the event that a given station meets all four criteria, i.e. belongs to our preferred clusters for the four variables relevant to this study.
  - C is actually the conjunction of all four of the R, D, H, S events "the station belongs to a selected cluster for variable X", i.e. $C = R \cap D \cap H \cap S$ (See Table 2).
- A is the event that a given station appears in our shortlist.
- T_i is the event that a given station is of type T_i (See Table 1).
  - For example, if M_X is the event "the station measured variable X", then the event T_7 is equivalent to $M_R \cap \neg M_D \cap M_H \cap M_S$
- h is the event that a given station is located at an elevation h.

Thus, the probability that a station $j = \{0, 1, ..., N_{\mathrm{shortlist}}\}$ of type T_i, meets all four criteria (C) and appears in our shortlist (A) given that it is located at an elevation h, can be written as $P_j = P(C \cap A \cap T_i \mid h)$. We cannot calculate exactly this probability, as it depends on unknown unknowns, i.e. on whether a station that did not measure a given variable in reality belongs to one of our preferred climatological clusters for that variable (Table 2). Thus we will

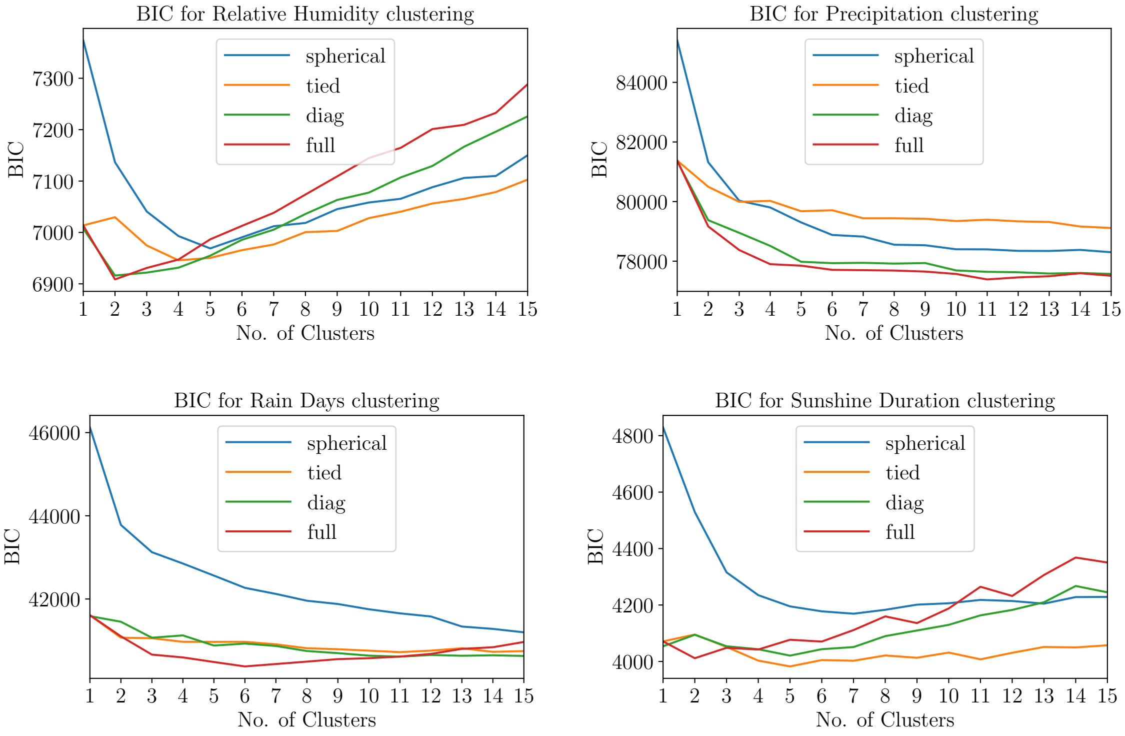

Figure 3. Bayesian Information Criterion (BIC) results for Gaussian Mixture Models using a different number of clusters and using different covariance restriction methods (Pedregosa et al. 2011) applied on each measured variable dataset.

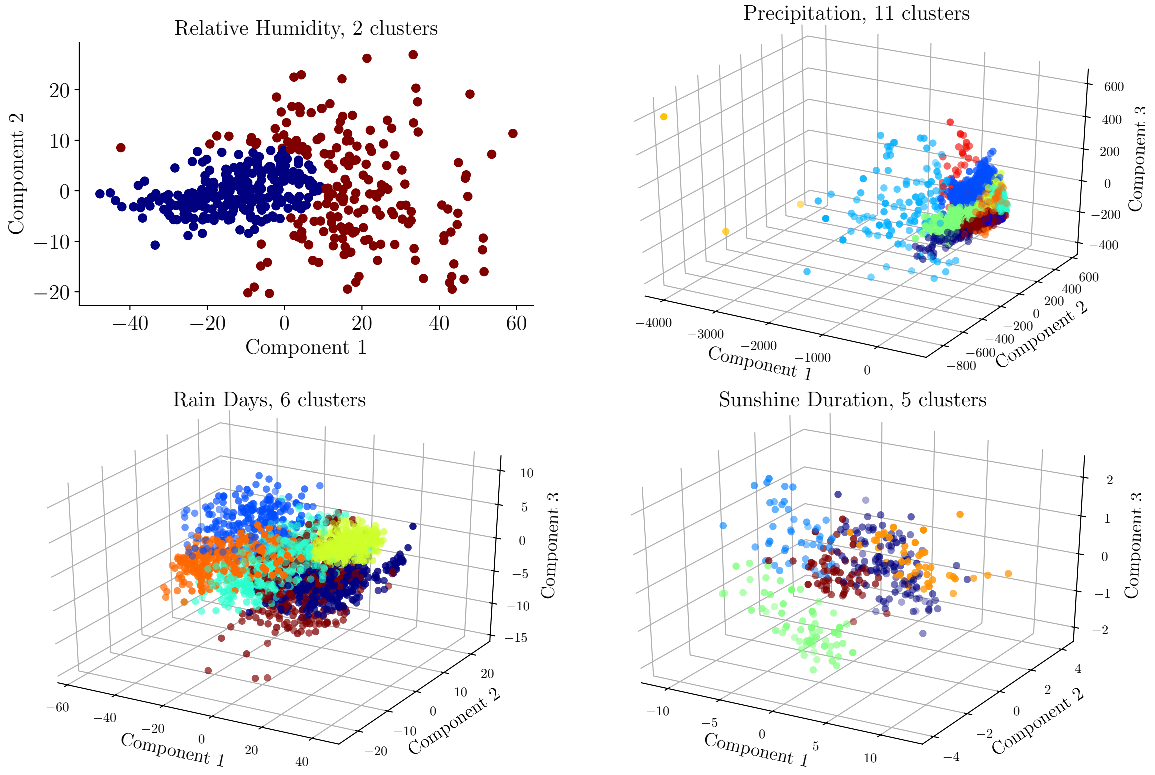

Figure 4. Lowest BIC Gaussian Mixture Model clustering of reduced dimensionality climate data. Colors indicate the results of the clustering classification.

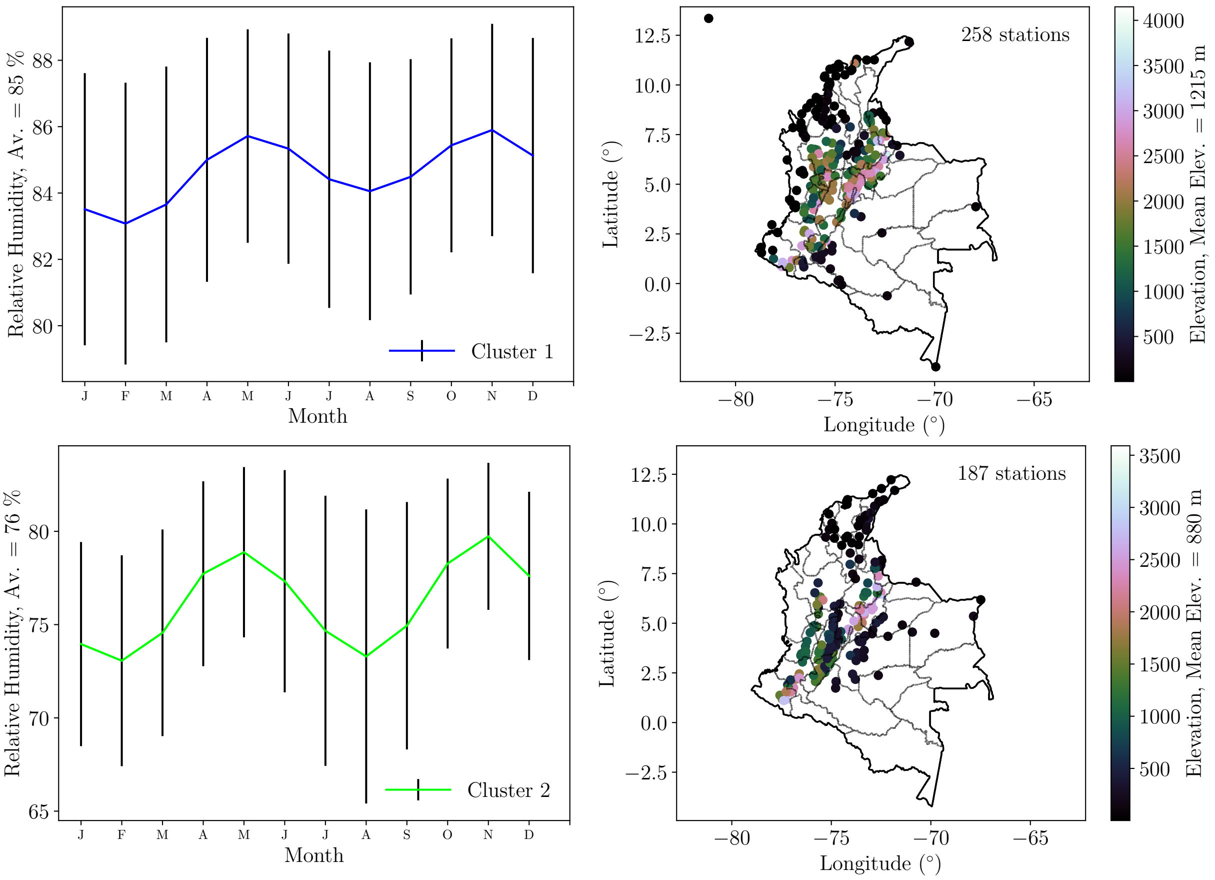

Figure 5. Climate clusters (1-2 of 2) as predicted by the lowest-BIC Gaussian Mixture Model for relative humidity. Left column: Average monthly relative humidity and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each relative humidity cluster.

use Bayes' theorem to help estimate this probability using simpler conditional probabilities. First we control for the inhomogeneous distribution of elevations $P(h)$,

$$P_j = P(C \cap A \cap T_i \mid h) = \frac{P(C \cap A \cap T_i \cap h)}{P(h)} \ . \quad (1)$$

The joint probability of all events $C \cap A \cap T_i \cap h$ can be rewritten as,

$$P_j = \frac{P(h \mid C \cap A \cap T_i)}{P(h)} P(C \cap A \cap T_i) \ . \quad (2)$$

If a station satisfies all four criteria, $P(C \cap A \cap T_i) = 1$. This probability can be expressed in terms of simpler conditional probabilities,

$$P_j = \frac{P(h \mid C \cap A \cap T_i)}{P(h)} P(C \mid A \cap T_i) P(A \mid T_i) P(T_i) \ , \quad (3)$$

where

- $P(h \mid C \cap A \cap T_i)$ is the distribution of elevations for a station type (T_i) on our shortlist ($C \cap A$),
- $P(A \mid T_i)$ is the probability that a station of type T_i is in our shortlist,
- $P(T_i)$ is the probability that a station type is T_i, and
- $P(C \mid A \cap T_i)$ is the probability that a station meets all four criteria (C) given that it is in our shortlist (A) and is of type T_i.

All of these probabilities except for the last one can be computed directly from the data[1].

The probability $P(C \mid A \cap T_i)$ has to be estimated indirectly, as we cannot presume to know under which conditions a station can satisfy all our criteria. However, we can rewrite this probability as,

$$P(C \mid A \cap T_i) = \frac{P(C \cap A \cap T_i)}{P(A \cap T_i)} \ . \quad (4)$$

In order to estimate $P(C \mid A \cap T_i)$ we assume that the probability of a station meeting our criteria for a given number of variables is independent of the station type. To illustrate this, let us assume that a given station is of type 1, so it only measured precipitation and rain days, i.e.,

$$T_1 = M_R \cap M_D \cap \neg M_H \cap \neg M_S \ . \quad (5)$$

If this station of type T_1 is in our shortlist (A), the R and D criteria are already met, so $A \cap R \cap D = A$. Thus,

$$P(C \cap A \cap T_1) = P(H \cap S \mid A \cap T_1) \ . \quad (6)$$

[1] Distributions of elevation-related probabilities were approximated discretely using a bin size estimated from the mean elevation difference between stations when sorted by elevation.

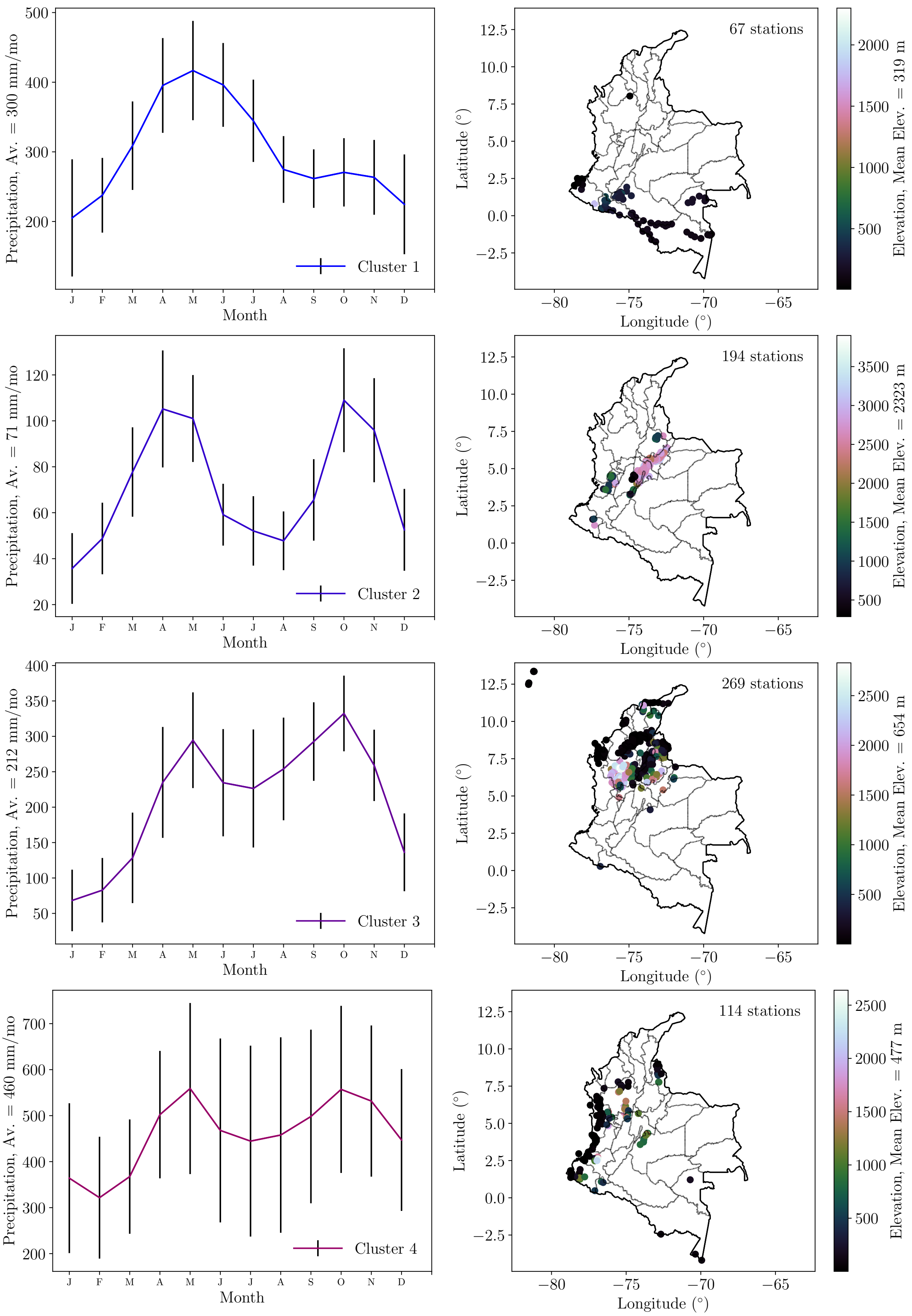

Figure 6. Climate clusters (1-4 of 11) as predicted by the lowest-BIC Gaussian Mixture Model for precipitation. Left column: Average monthly precipitation and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each precipitation cluster.

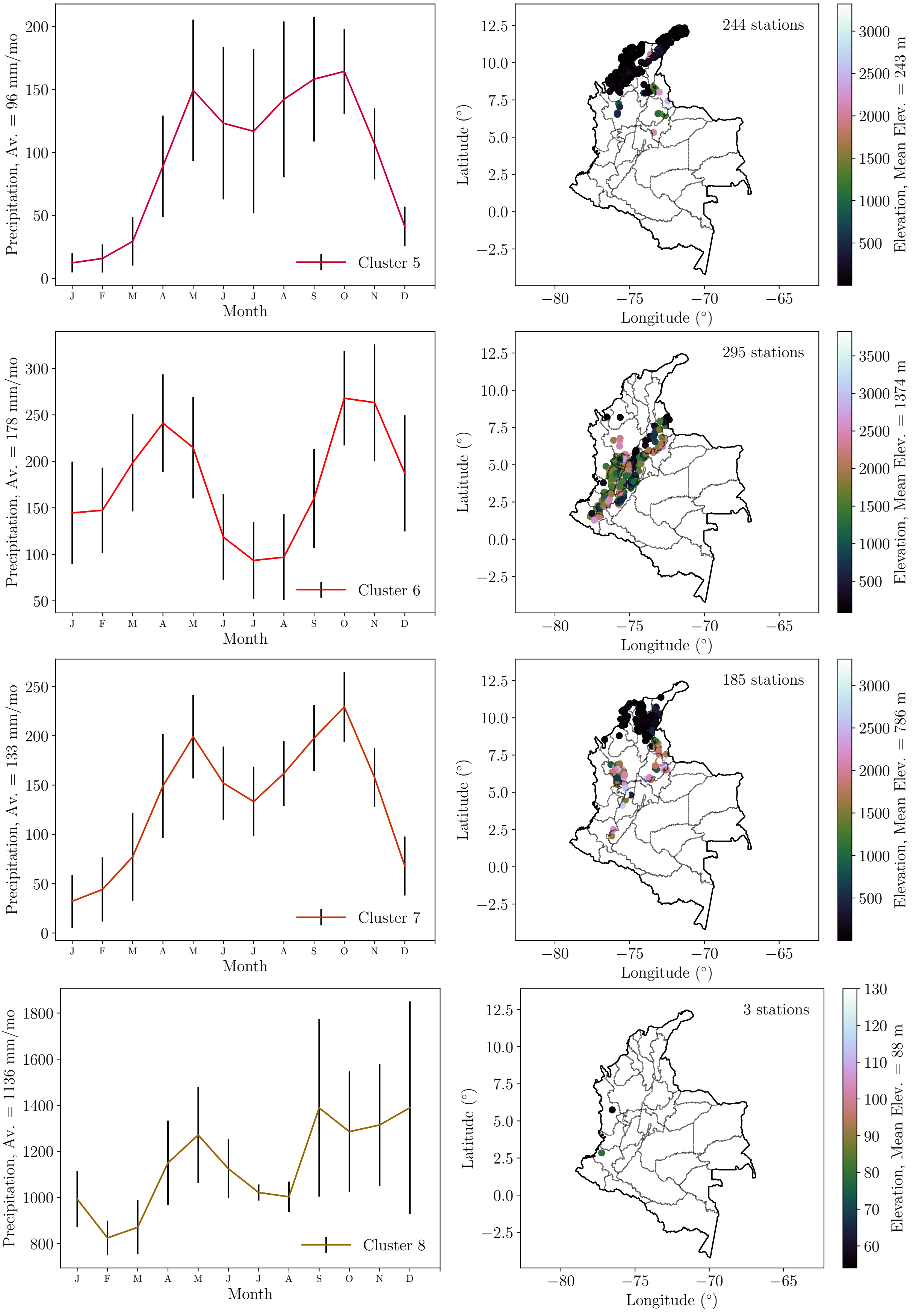

Figure 7. Climate clusters (5-8 of 11) as predicted by the lowest-BIC Gaussian Mixture Model for precipitation. Left column: Average monthly precipitation and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each precipitation cluster. Continuation of Figure 6.

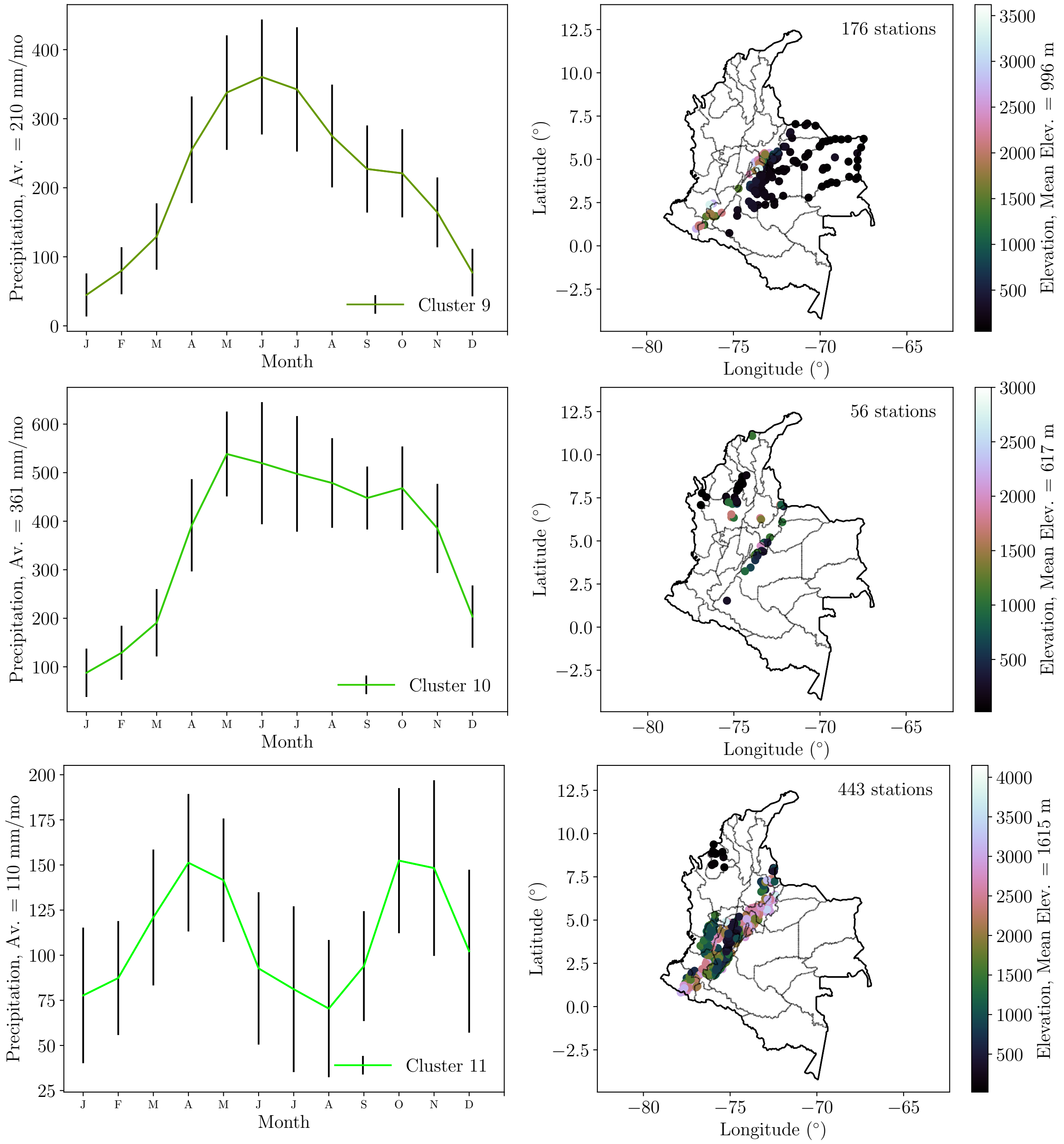

Figure 8. Climate clusters (9-11 of 11) as predicted by the lowest-BIC Gaussian Mixture Model for precipitation. Left column: Average monthly precipitation and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each precipitation cluster. Continuation of Figure 7.

Substituting this into Eq. (4),

$$P(C \mid A \cap T_1) = \frac{P(H \cap S \mid A \cap T_1)}{P(A \cap T_1)} = P(H \cap S \mid A \cap T_1)\,. \quad (7)$$

The last term in the previous expression is unknown, but we can approximate it by assuming that it is independent of the station type. Therefore,

$$P(C \mid A \cap T_1) \simeq P(H \cap S)\,. \quad (8)$$

The probability on the right hand side of the previous equation can be directly computed from the data, as it only requires counting how many stations meet both H and S criteria. This procedure can be extended to all other station types.

The probabilities for stations in our shortlist (P_j in Equation 3) can be used to evaluate the quality of a site, but since they vary by orders of magnitude, for the sake of clarity we decided to use instead a probability-derived logarithmic quality index. Thus, we define the quality index Q_j for a station j in our shortlist as,

$$Q_j = 9\,\frac{\log(^{P_{\rm j}}/_{P_{\rm min}})}{\log(^{P_{\rm max}}/_{P_{\rm min}})} + 1\,. \quad (9)$$

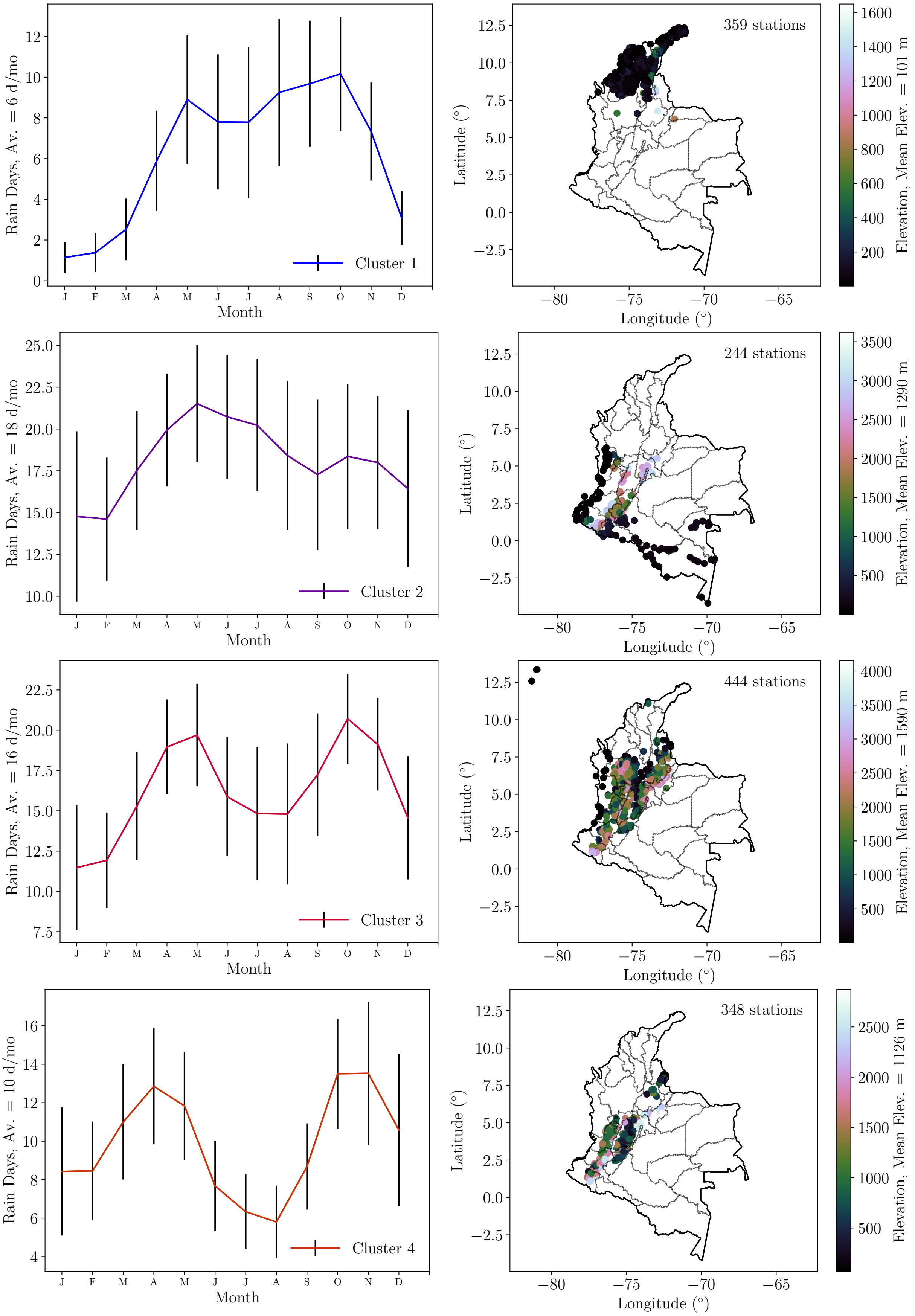

Figure 9. Climate clusters (1-4 of 6) as predicted by the lowest-BIC Gaussian Mixture Model for rain days. Left column: Average monthly rain days and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each rain days cluster.

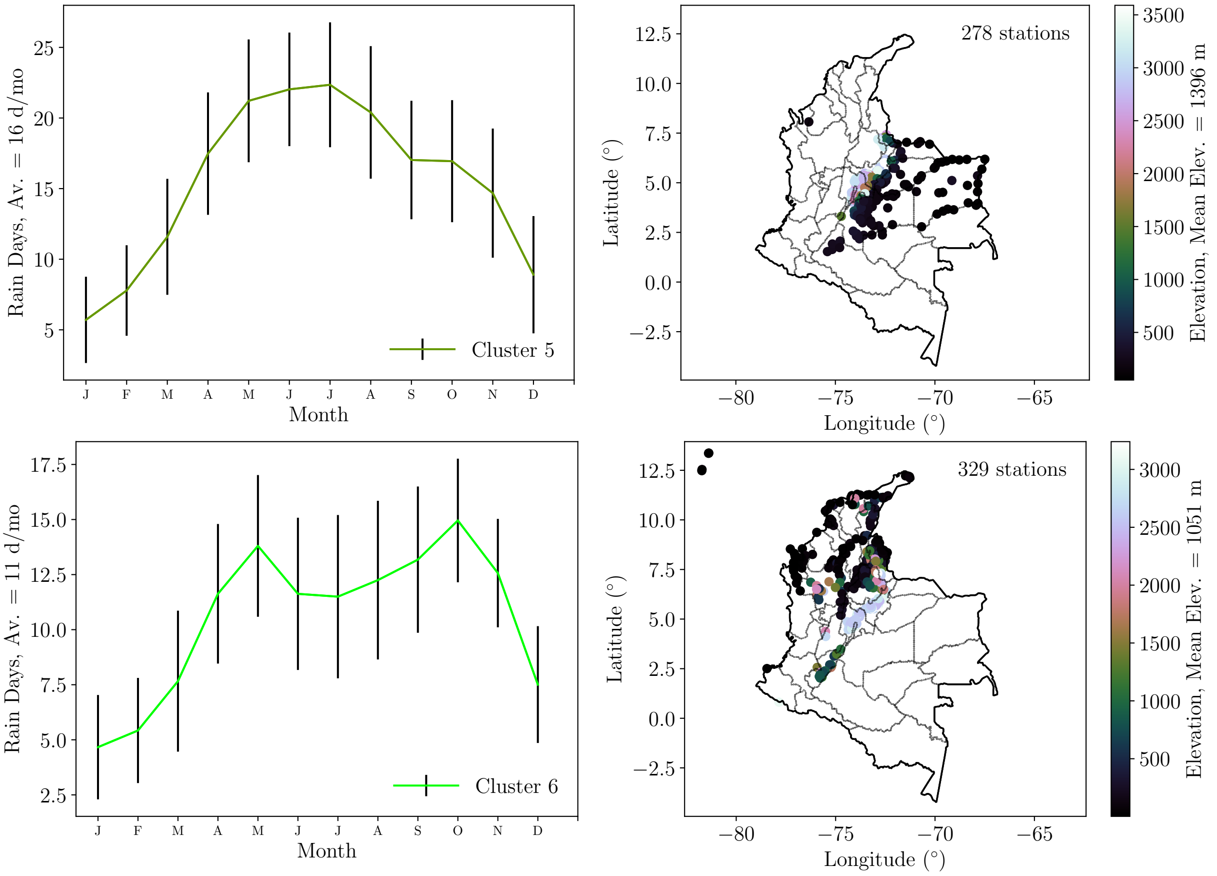

Figure 10. Climate clusters (5-6 of 6) as predicted by the lowest-BIC Gaussian Mixture Model for rain days. Left column: Average monthly rain days and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each rain days cluster. Continuation of Figure 9.

Here $P_{\min}, P_{\max}$ are the lowest and highest values for the probabilities for the stations in the first shorlist of 665 stations that satisfy criteria for all measured variables. Thus, if $P_j = P_{\max}$, $Q_j = 10$, and if $P_j = P_{\min}$, $Q_j = 1$.

4.2 Final selection

Many of our 83 candidate stations, selected using the method described at the beginning of this section, were lacking in relative humidity and sunshine duration data. Given that for stations less than 5 km away the difference in relative humidity and sunshine duration data is less than 15%, we extrapolated using data from nearby stations (at less than 4 km away and at an elevation difference of less than 100 m). The extrapolated data evidenced that 4 of the stations in the shortlist did not satisfy the relative humidity and sunshine duration criteria. The remaining 79 stations are mapped in Figure 13a.

We rejected 9 of the stations in Figure 13a, i.e. those located outside of the region located within the latitude, longitude range $[(4.3, 6.2)°, (-72.4, -74.5)°]$. Even though they could be indicative of unusually dry, sunny regions, there are simply not enough stations nearby to say anything conclusively, as evidenced in Figure 14. Adding to this, the only variable they measured was precipitation ($Q_j = 2.2$), which leads us to reject them from our shortlist.

The final 70 stations located in the Cundinamarca-Boyacá region of Colombia are shown in more detail in Figure 13b. This figure shows two candidate locations in the northern region (Boyacá department), where the quality index Q_j (proportional to the point size) indicates that all four criteria are satisfied. In order to see if those regions are geographically correlated, we grouped together these stations in a simplified manner using a lowest-BIC spherical covariance Gaussian Mixture Model on their coordinates (Figure 15).

5 RESULTS

We selected clusters of climatologically similar stations across Colombia using an unsupervised learning low-dimensional embedding algorithm for each measured variable. From these clusters, 70 weather stations came up as potential candidates for identifying high-mountain regions (at an elevation greater that 2000 masl) with an unusually dry, sunny weather. Seasonally, the driest months of the year for all the candidate regions of interest are December-February and June-August (Figures 5-12).

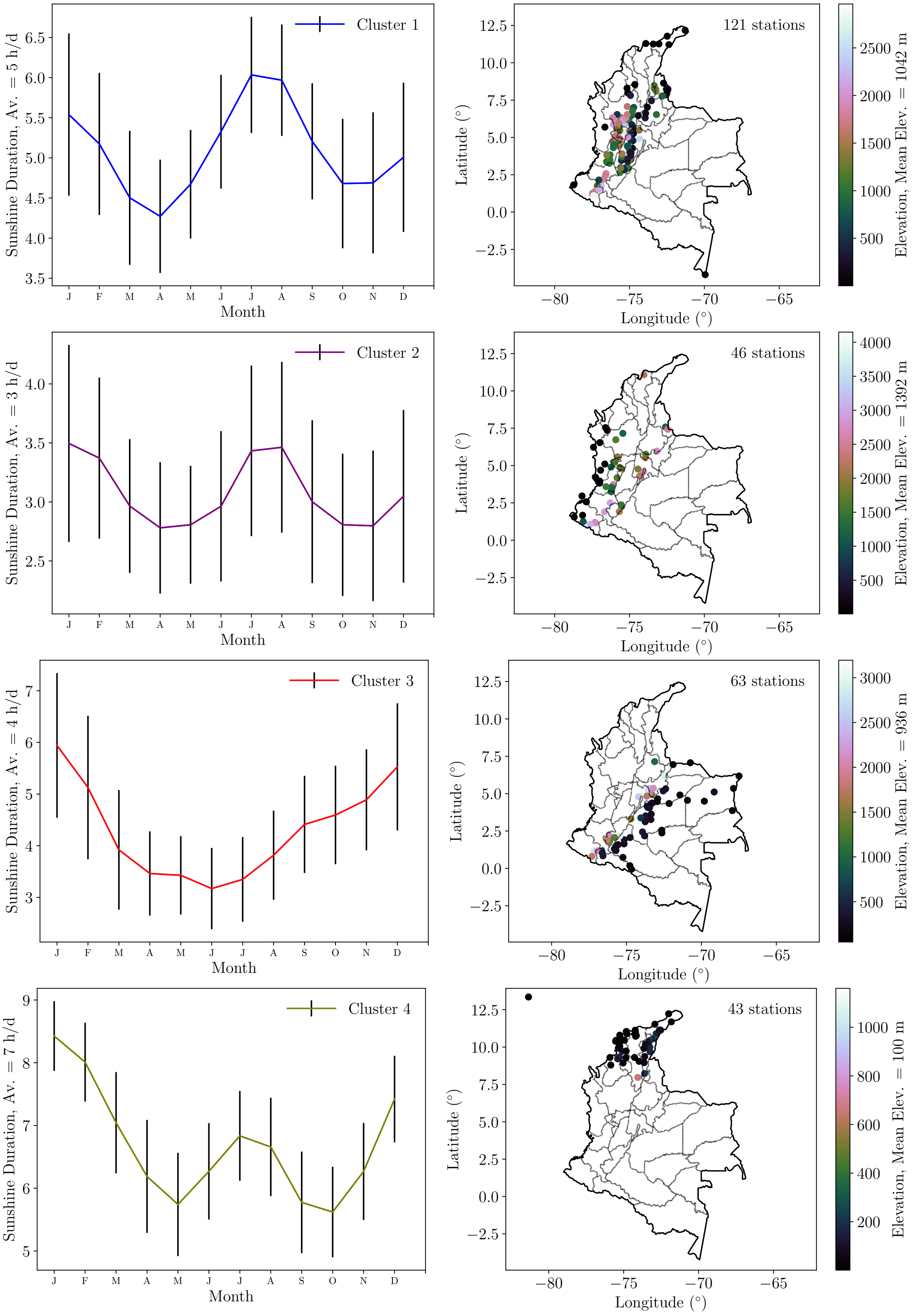

Figure 11. Climate clusters (1-4 of 5) as predicted by the lowest-BIC Gaussian Mixture Model for sunshine duration. Left column: Average monthly sunshine duration and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each sunshine duration cluster.

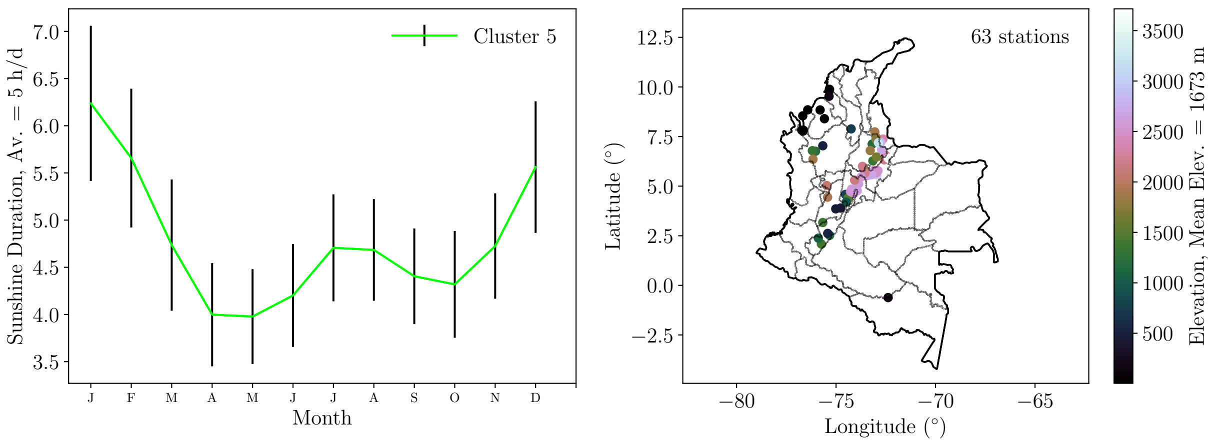

Figure 12. Climate clusters (5 of 5) as predicted by the lowest-BIC Gaussian Mixture Model for sunshine duration. Left column: Average monthly sunshine duration and standard deviation (error bars) for each climate cluster. Right column: Geographical location and elevation of weather stations in each sunshine duration cluster. Continuation of Figure 11.

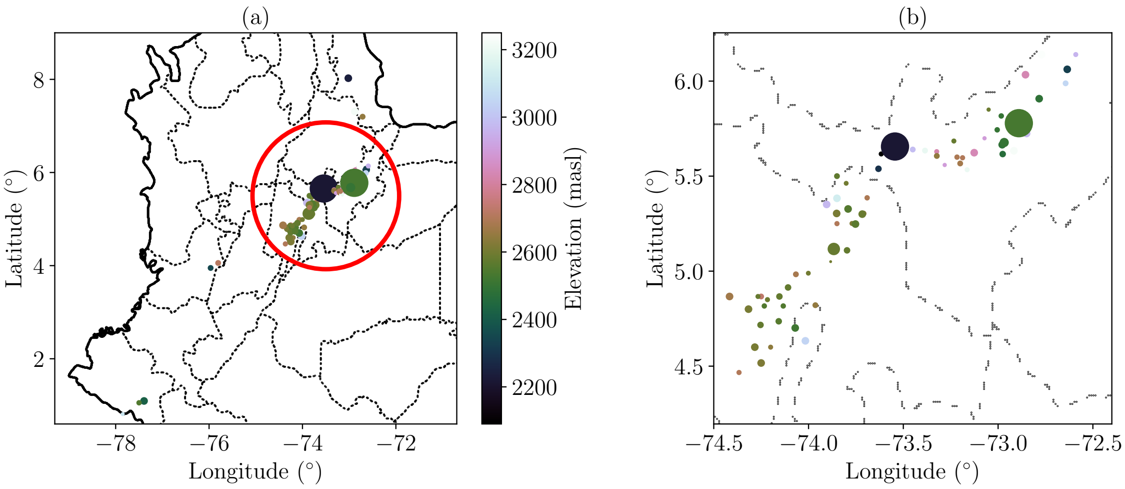

Figure 13. Map of stations in our shortlist. The highlighted region in (a) is seen in more detail in (b), which corresponds approximately to the Colombian *altiplano cundiboyacense* region. Marker size is proportional to the probability P_j, where the largest size corresponds to $Q_j = 10$.

5.1 Candidate regions of interest

Figure 15 shows 6 candidate regions of interest, where weather stations indicating dry, sunny climate are regionally correlated within a $10 - 20$ km radius. In order to visualize the extent of those regions we plotted the predicted variance regions for each cluster along with other (rejected) stations in our sample in Figure 15. It should be noted that the spherical variance GMM geographical clustering overestimates the extent of each region, and tells us nothing of their actual geographical shape.

We can now reject regions where the number of rejected stations within the predicted variance regions is similar or greater than the number of stations in our shortlist in Figure 15. We discuss the suitability of each of these regions below, from South-West to North-East based on our Bayesian probabilistic quality index (Section 4.1).

5.1.1 La Sabana

The first cluster is located at (4.77°N, 74.18°W) and covers an radius of 15.3 km (Figure 15, purple circle). This is a region located to the west of Bogotá at an elevation above 2600 masl, with a large variety of microclimates. This explains the significant amount of stations from this region in our shortlist. However, Figure 15 shows that within this predicted variance region there are more stations that do not satisfy our criteria than stations that do, and no single sub-region can be identified. This, compounded with the fact that most (70%) of these stations have a Q value lower than 5.6 due to the lack of humidity/sunshine data, implies that at the moment this is not a strong candidate region of interest.

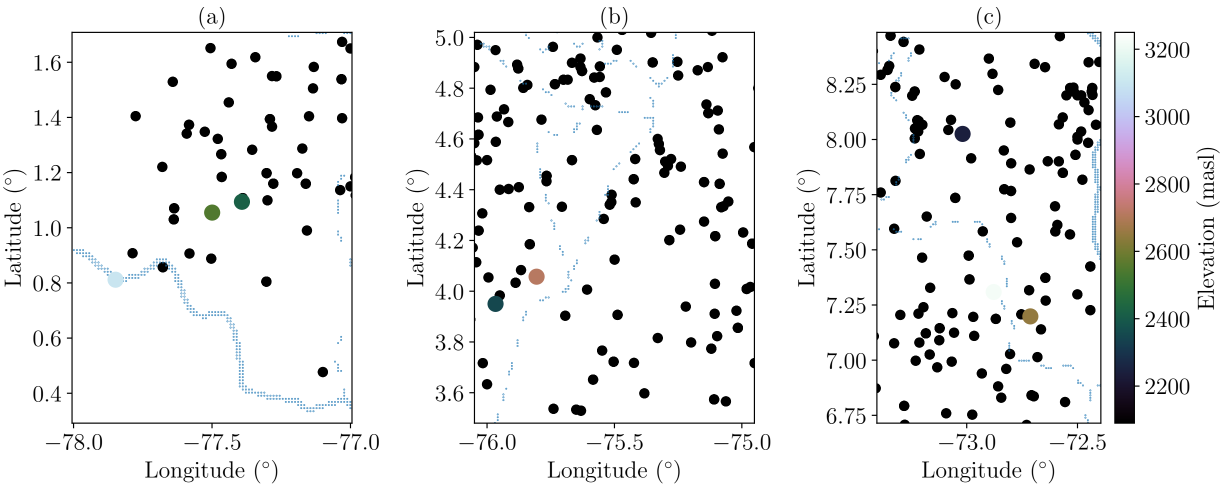

Figure 14. Map of stations in our final shortlist but outside of the Colombian *altiplano* region highlighted in Figure 13. Marker size is proportional to the probability P_j, where the smallest size corresponds to $Q_j = 2.2$.

5.1.2 Valle de Ubaté

This cluster is located at (5.30°N, 73.80°W) and covers a radius of 11.4 km (Figure 15, blue circle). This region, located to the NNE of Bogotá is a hilly region along a valley at an elevation above 2600 masl. The ratio of number of stations that do not meet our criteria to the stations in this predicted variance region is low (0.3), and almost half of these stations (46%) have a Q_j value higher than 6. This indicates a potential candidate region, although more sunshine and humidity data are needed.

5.1.3 Villa de Leyva

This cluster is located at (5.61°N, 73.55°W) and covers a radius of 7.11 km (Figure 15, yellow circle). This region is located near the town of Villa de Leyva, and it is known for its dry weather. The ratio of stations not in our shortlist vs. the stations in this predicted variance region is not very low (0.4). However, Figure 15 seems to indicate that the actual region of interest is narrower than the predicted variance region, as the stations that do not satisfy our criteria are on the NNW and SSW fringes of said region. The presence of one very high-Q_j station and the narrow GMM-predicted variance is indicative of this region being a strong candidate. However, its relatively low elevation (near 2200 masl) can signify the presence of too much atmospheric water vapor above the surface.

5.1.4 Cantón de Tunja

This cluster is located at (5.59°N, 73.24°W) and covers a radius of 8.29 km (Figure 15, red circle). The GMM-predicted variance region west of Tunja is located in the historic Tunja Canton, and even though the ratio of stations not in our shortlist to stations in this predicted region is low (0.31), there are simply not enough humidity and sunshine data to make any strong conclusion about this region. However, the high elevation of some of these stations is tantalizing, which is why we will keep this as a region of interest for our next paper.

5.1.5 Valle del Sol

This cluster is located at (5.72°N, 72.96°W) and covers a radius of 8.06 km (Figure 15, green circle). This is one of the most promising regions in our sample. Located in a wide, sunny valley (hence the name) at an elevation of 2600 masl, it is surrounded by mountains, and rains are not as common as elsewhere in the country. This region has the lowest ratio of not in our shortlist stations to stations in this predicted variance region (0.2), 40% of the stations have a Q_j value higher than 6, and one of the stations has a very high-Q_j value. This will be one of the regions of interest for our next paper, and we think that it warrants radiometer measurements to be carried out.

5.1.6 Parque Nacional Natural Pisba

This cluster is located at (6.04°N, 72.71°W) and covers a radius of 10.52 km (Figure 15, dark green circle). This is a *páramo* region, characterized by high-mountain tundra weather, with some mountains reaching an elevation of up to 3800 masl. The ratio of stations rejected stations to stations in this predicted variance region is high (0.5), and it is unlikely that the mist allows for a low-atmospheric water vapor region to be located here, even if rain is sporadic. The lack of a high-Q station means that more sunshine-humidity data are needed, but this is possibly not a region of interest.

5.2 Comparison to other works

We compared our results to the work of Suen et al. (2014) and Suen (2016), where they obtained global precipitable water vapor maps for 2012 using data from the MODIS instrument on board the Aqua and Terra satellites. The lowest mean precipitable water vapor (∼ 15 mm) regions in the Suen et al. (2014) map for Colombia are near the Ubaté, Tunja, Valle del Sol, and Pisba regions. The Suen (2016) global median precipitable water vapor map has an upper bound of 10 mm for stations located in the clusters corresponding to those same regions. Unfortunately, the resolu-

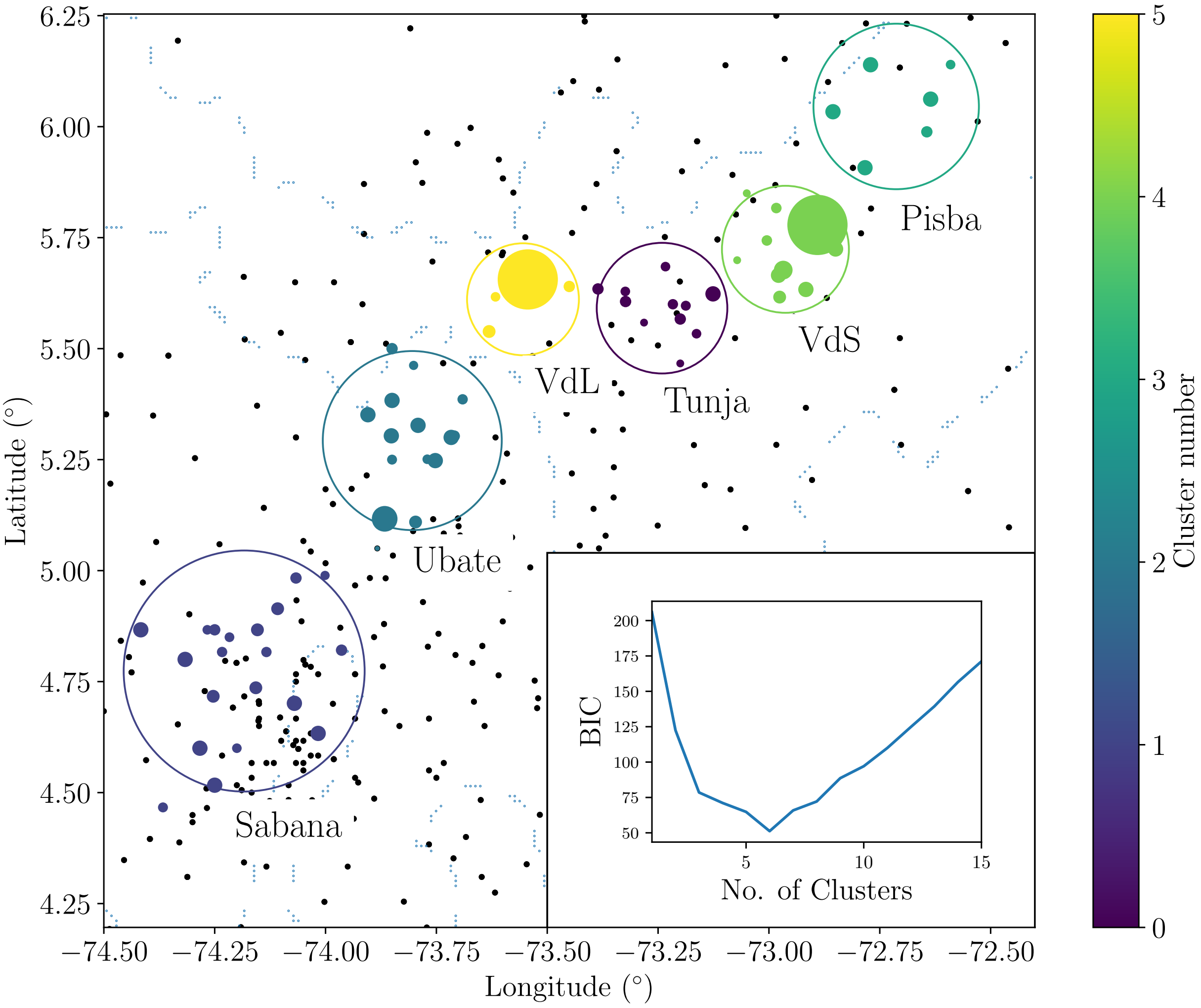

Figure 15. Geographical clustering results for a lowest-BIC spherical covariance Gaussian Mixture Model of the stations located inside the red circle in Figure 13b, indicating the predicted variance regions for each cluster (red circles), and plotted along rejected stations in our original dataset (black points). Marker size is proportional to the probability P_j, where the largest size corresponds to $Q_j = 10$ and the smallest size corresponds to $Q_j = 2.2$. Embedded plot: BIC results for different number of clusters.

tion of the maps is not fine enough to pinpoint a site location at a precision higher than ∼ 20 km.

6 CONCLUSIONS

In this paper we find plausible locations for a high-mountain mm-wave astronomical observatory in the northern Andes of Colombia, by analizing 30 years of climate data from 2046 weather stations. By low-dimensionally embedding the data, we are able to group together and correlate climate behavior indicative of daytime dry, clear-sky weather conditions. A repository with all our code and the full dataset is available at `https://github.com/saint-germain/ideam`.

From our shortlist of 79 stations at elevations higher than 2000 masl that met our criteria, we selected 70 stations located in the Cundinamarca-Boyacá *altiplano* region of Colombia, which is a high-mountain basin with an average elevation of 2600 masl. Weather stations in our shortlist but outside this region are too sparsely located to suggest a candidate region of interest.

From those 70 stations we obtained 6 geographically correlated candidate regions of interest: La Sabana, Valle de Ubaté, Cantón de Tunja, Villa de Leyva, Valle del Sol, and Parque Nacional Natural Pisba. Seasonally, there appear to be two times of the year (Dec-Feb and Jun-Aug) where the weather conditions indicate the best daytime dry, clear-sky conditions. These months could be the best for mm-wave astronomical observations to be carried out.

We summarize the information about two regions of interest (Valle de Ubaté and Valle del Sol) in Table 3, selected based on a Bayesian probabilistic quality index. Most of the stations in this table are located at elevations

Table 3. Stations in two regions of interest showing a high-elevation daytime dry, clear-sky climate, organized by location, Bayesian probability quality measure (Q_j), distance to cluster center (r_c), mean precipitation ($\bar{R}$), mean rain days ($\bar{D}$), mean relative humidity ($\bar{H}$), mean sunshine duration ($\bar{S}$), and elevation (h). Detailed table (including all clusters) available at `https://github.com/saint-germain/ideam` .

Region of interest	Station code	Municipality	Q_j	r_c (km)	$\bar{R}$ (mm/mo)	$\bar{D}$ (d/mo)	$\bar{H}$ (%)	$\bar{S}$ (h/d)	h (masl)
Valle de Ubaté	21201410	Nemocon	2.2	28	66	ND	ND	ND	2600
	24011060	Susa	4.6	18	83	12	ND	ND	2600
	24010140	Cucunuba	4.7	6	57	9	ND	ND	2620
	24015210	Sutatausa	4.8	6	57	10	ND	ND	2700
	24010170	Guacheta	4.9	16	73	10	ND	ND	2690
	24010070	Lenguazaque	5.2	10	63	12	ND	ND	2650
	24015130	Simijaca	5.2	23	67	11	ND	ND	2572
	21201620	Suesca	5.6	20	59	10	ND	ND	2575
	24010610	Carmen de Carupa	6.1	12	54	12	ND	ND	2970
	24011080	Cucunuba	6.1	7	51	5	ND	ND	2562
	24010280	Lenguazaque	6.1	9	57	10	ND	ND	2585
	24010440	Susa	6.1	11	60	8	ND	ND	3130
	24011090	Ubate	6.1	4	62	13	ND	ND	2555
	24015110	Ubate	6.1	5	62	12	ND	ND	2610
	21205400	Nemocon	7.6	20	52	10	77	4	2580
Valle del Sol	24030760	Duitama	4.0	17	69	10	ND	ND	2590
	24030510	Paipa	4.0	12	74	9	ND	ND	2900
	24030790	Nobsa	5.0	6	66	10	ND	ND	2500
	24031040	Santa Rosa	5.0	10	70	9	ND	ND	2500
	24035140	Sogamoso	5.0	4	59	10	ND	ND	2500
	24030940	Sogamoso	5.0	7	60	12	ND	ND	2500
	24030410	Tibasosa	5.0	5	62	7	ND	ND	2500
	24030410	Iza	5.6	12	52	9	ND	ND	2470
	24030540	Firavitoba	6.1	6	57	9	ND	ND	2486
	24030190	Mongui	6.1	12	65	9	ND	ND	2970
	24030760	Sogamoso	6.1	11	65	10	ND	ND	3225
	24035340	Sogamoso	6.7	5	61	12	75	5	2500
	24035150	Nobsa	10.0	10	68	13	74	4	2530

above 2600 masl and below 3000 masl. We validated our results using satellite measurements of the upper bound for mean precipitable water vapor (Suen et al. 2014; Suen 2016) for those two regions, which is reported to be $\sim$ 15 mm in $\sim$ 20 km resolution maps. This means that it is not implausible to find a site with an even lower water vapor column in one of the regions of interest identified here. Some mountains nearby Valle de Ubaté and Valle del Sol have elevations of up to 3400 masl, and therefore warrant further measurements using radio sondes and radiometers.

The Villa de Leyva region, despite having a dry, sunny climate, is located at a comparatively low elevation (2200 masl), which means that the precipitable water vapor column above this region is too high ($>$ 20 mm according to the Suen (2016) maps).

There is an additional issue with the Valle del Sol region. The high industrial activity in this region fills the air with particulate material (Pizarro 2004; Jimenez Pizarro et al. 2011), which would nullify the suitability of this region. Besides obtaining seasonal high-resolution precipitable water vapor maps, in our next paper we will study the amount of particulate material in the reported regions of interest using MODIS (Aqua and Terra) satellite data.

The methods used here can be extended and adapted to similar climatological datasets in other regions. This can be a preliminary step in radio astronomy site testing in developing countries.

ACKNOWLEDGMENTS

This work was made possible by the Vicerrectoría de Investigación 01-2015 research grant from Universidad ECCI. We would also like to thank the Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) for providing us with the dataset.

REFERENCES

Archibald E. N., et al., 2002, Monthly Notices of the Royal Astronomical Society, 336, 1

Aumann H. H., et al., 2003, IEEE Transactions on Geoscience and Remote Sensing, 41, 253

Battistelli E. S., et al., 2012, Monthly Notices of the Royal Astronomical Society, 423, 1293

Bustos R., Rubio M., OtÃąrola A., Nagar N., 2014, Publications of the Astronomical Society of the Pacific, 126, 1126

Cadeddu M. P., Liljegren J. C., Turner D. D., 2013, Atmospheric Measurement Techniques, 6, 2359

Caumont O., et al., 2016, Quarterly Journal of the Royal Meteorological Society, 142, 2692

Chamberlin R. A., Grossman E. N., 2012, Journal of Geophysical Research: Atmospheres, 117, n/a

Cimini D., Westwater E. R., Gasiewski A. J., Klein M., Leuski V. Y., Liljegren J. C., 2007, IEEE Transactions on Geoscience and Remote Sensing, 45, 2169

Clough S. A., Kneizys F. X., Davies R. W., 1989, Atmospheric Research, 23, 229

Cortés F., Reeves R., Bustos R., 2016, Radio Science, 51, 1166

Denny S., Suen J., Lubin P., 2013, New Astronomy, 25, 114

He J., Zhang S., Wang Z., 2012, Radio Science, 47

Hosako I., et al., 2007, Proceedings of the IEEE, 95, 1611

Jimenez Pizarro R., Arango C. D., Peña J. A., 2011, AGU Fall Meeting Abstracts,

Jones A., et al., 2012, Atmospheric Chemistry and Physics, 12, 5207

Jones E., Oliphant T., Peterson P., et al., 2001–, SciPy: Open source scientific tools for Python, `http://www.scipy.org/`

Kuhn T., Bauer A., Godon M., BÃijhler S., KÃijnzi K., 2002, Journal of Quantitative Spectroscopy and Radiative Transfer, 74, 545

Lew B., Uscka-Kowalkowska J., 2016, Monthly Notices of the Royal Astronomical Society, 455, 2901

Liebe H. J., 1989, International Journal of Infrared and millimeter waves, 10, 631

Liljegren J. C., Clothiaux E. E., Mace G. G., Kato S., Dong X., 2001, Journal of Geophysical Research: Atmospheres, 106, 14485

Luini L., Riva C. G., 2016, IEEE Transactions on Antennas and Propagation, 64, 2487

Niell A. E., Coster A. J., Solheim F. S., Mendes V. B., Toor P. C., Langley R. B., Upham C. A., 2001, Journal of Atmospheric and Oceanic Technology, 18, 830

Paine S., Blundell R., Papa D. C., Barrett J. W., Radford S. J., 2000, Publications of the Astronomical Society of the Pacific, 112, 108

Pardo J. R., Cernicharo J., Serabyn E., 2001a, IEEE Transactions on Antennas and Propagation, 49, 1683

Pardo J., Serabyn E., Cernicharo J., 2001b, Journal of Quantitative Spectroscopy and Radiative Transfer, 68, 419

Pazmany A. L., 2007, IEEE Transactions on Geoscience and Remote Sensing, 45, 2202

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Peng S. S., Wu L., Ying X. H., Xu Z. C., 2009, Journal of Infrared, Millimeter, and Terahertz Waves, 30, 259

Peter R., Kämpfer N., 1992, Journal of Geophysical Research: Atmospheres, 97, 18173

Pickett H. M., Poynter R. L., Cohen E. A., Delitsky M. L., Pearson J. C., Müller H. S. P., 1998, Journal of Quantitative Spectroscopy and Radiative Transfer, 60, 883

Pinzón G., González D., Hernández J., 2015, Publications of the Astronomical Society of the Pacific, 127, 523

Pizarro R. J., 2004, PhD thesis, École Polytechnique Fédérale de Lausanne

Radford S. J. E., Peterson J. B., 2016, Publications of the Astronomical Society of the Pacific, 128, 075001

Reynolds D., 2009, Gaussian Mixture Models. Springer US, Boston, MA, pp 659–663, doi:10.1007/978-0-387-73003-5_196, `"http://dx.doi.org/10.1007/978-0-387-73003-5_196"`

Rosenkranz P. W., 1998, Radio Science, 33, 919

Schwarz G., 1978, Ann. Statist., 6, 461

Seidel D. J., Fu Q., Randel W. J., Reichler T. J., 2008, Nature Geoscience, 1, 21

Shaw B., Jebara T., 2009, in Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09. ACM, New York, NY, USA, pp 937–944, doi:10.1145/1553374.1553494, `http://doi.acm.org/10.1145/1553374.1553494`

Slocum D. M., Slingerland E. J., Giles R. H., Goyette T. M., 2013, Journal of Quantitative Spectroscopy and Radiative Transfer, 127, 49

Smith G. J., Naylor D. A., Feldman P. A., 2001, International Journal of Infrared and Millimeter Waves, 22, 661

Suen J. Y., 2016, Journal of Infrared, Millimeter, and Terahertz Waves, 37, 615

Suen J. Y., Fang M. T., Lubin P. M., 2014, IEEE Transactions on Terahertz Science and Technology, 4, 86

Turner D. D., Cadeddu M. P., Lohnert U., Crewell S., Vogelmann A. M., 2009, IEEE Transactions on Geoscience and Remote Sensing, 47, 3326

Wang J., Zhang L., Dai A., Van Hove T., Van Baelen J., 2007, Journal of Geophysical Research: Atmospheres, 112, n/a

Wentz F. J., Meissner T., 2016, Radio Science, 51, 381

Withayachumnankul W., Fischer B. M., Abbott D., 2008, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 464, 2435

This paper has been typeset from a TEX/LATEX file prepared by the author.