Site location study for a high-mountain millimeter observatory in the Colombian Andes I: In situ data

Germán Chaparro Molano[1]

[1]Grupo de Simulación, Análisis y Modelado, Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia

E-mail: gchaparrom@ecci.edu.co

Oscar Leonardo Ramírez Suárez[1]

[1]Grupo de Simulación, Análisis y Modelado, Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia

E-mail: oramirezs@ecci.edu.co

Oscar Restrepo[1]

[1]Grupo de Simulación, Análisis y Modelado, Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia

E-mail: orestrepog@ecci.edu.co

Alexander Martínez[1,2]

[1]Grupo de Simulación, Análisis y Modelado, Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia
[2]Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá, Colombia

December 2016

Abstract. This document describes the preparation of an article using LaTeX 2_ε and `iopart.cls` (the IOP Publishing LaTeX 2_ε preprint class file). This class file is designed to help authors produce preprints in a form suitable for submission to any of the journals listed in table **??** on the next page. You are not obliged to use this class file—we accept submissions using all common LaTeX class and style files. The `iopart.cls` class file is supplied merely as a convenience for those authors who find it useful. This document gives both general advice that applies whatever class file you use, and specific advice that applies if you choose to use `iopart.cls`.

We also accept submissions in Word format. Instructions for Word submissions are available via the 'Author Guidelines' link at `http://authors.iop.org`.

If you have any queries about this document or any aspect of preparing your article for submission please contact us at the e-mail address given above.

Submitted to: *PASP*

1. Introduction

Astronomical observations in the millimeter and sub-millimeter wavelength range require that atmospheric effects affecting absorption at these wavelengths are kept to a minimum. The main factor contributing to the atmospheric opacity is water vapor, which very efficiently absorbs light in the THz range due to BLAHBLAHBLAH. 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, satellite measurements, or indirectly via models of in situ climatological measurements or GPS-delay studies.

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. For this reason, the geographic characterization of the quality of a site is given in terms of the local water vapor column above the potential site. 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. 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.

You (Shaw & Jebara 2009)

2. Methods

The meteorological data used in this study was gathered for 30 years (1980-2010) in 2046 Instituto de Hidrología, Meteorologa y Estudios Ambientales (IDEAM) weather stations across Colombia (Figure **??**). The variables that we considered relevant to our work were: Precipitation (mm/mo), Rainy Days (d/mo), Relative Humidity (%), and Sunshine (h/d). These variables are reported as multi-annual monthly data, i.e. each variable is reported monthly from January to December averaged over the 1980-2010 range. For precipitation, each monthly datum corresponds to the cumulative monthly value. For rainy days, each monthly datum corresponds to the number of rainy days for that month. For relative humidity, each monthly datum corresponds to the average daily value averaged over each month. For sunshine, each monthly datum corresponds to the cumulative daily value averaged over each month. All stations reported precipitation values, although only 2002 reported rainy days values, 445 reported relative humidity values, and 336 reported sunshine values. This means that not all available stations registered all meteorological variables relevant to this study. We classify stations according

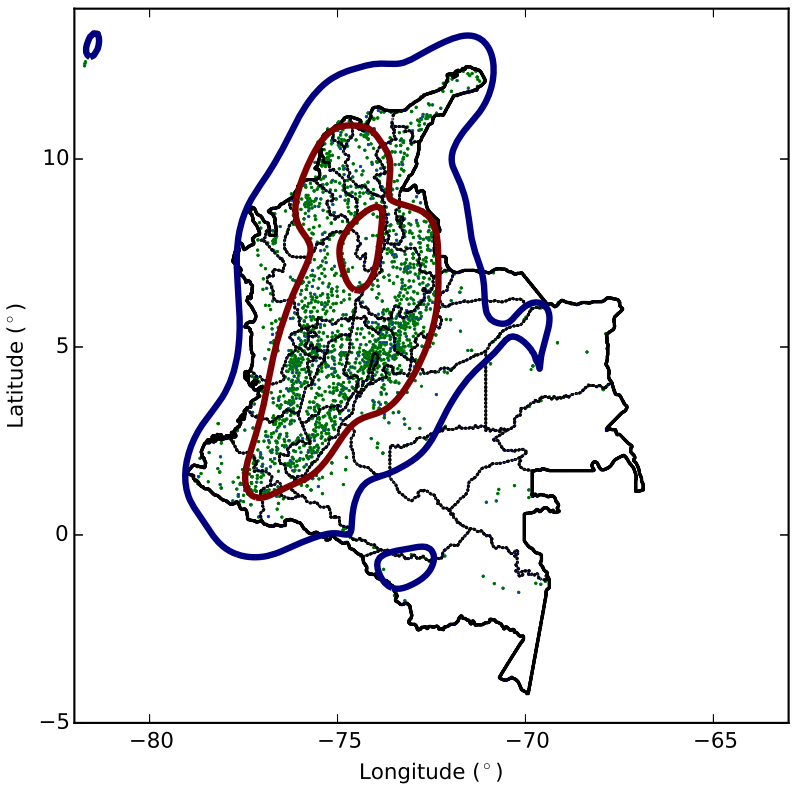

Figure 1. Say What

Table 1. Types of stations classified according to measured variables.

Type	Precip.	Rainy Days	Relative Humidity	Sunshine	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

to which variables they measured (Table 1).

Since the data represent multi-annual seasonal variations, we wish to classify and group together stations that show a similar climatological behavior for each variable, regardless of the station's location, elevation or type. This way, we can relate 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 present at high-mountain sites in the Andes mountain range.

We classified stations using a low-dimensional embedding (Shaw & Jebara 2009) for each variable, which ensures that a classification algorithm can uncover features in the data across a number of dimensions that is lower than the native dimensionality of the data (which in this case is monthly, i.e. $N_{\rm dim} = 12$). This involves 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 too many 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 yielding a low/locally minimum Bayesian Information Criterion (BIC) value.

2.1. 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, rainy days and sunshine, $N_{\rm dim}^{\rm PCA}(95\%) = 3$ 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).

2.2. Gaussian Mixture Models

After projecting the data along the components found by the PCA algorithm, we clustered the data using a Gaussian Mixture Model, which uses an Expectation-Maximization algorithm to find a Maximum Likelihood model made of a given number $N_{\rm G-C}$ of Gaussian distributions, each of which is expected to represent a cluster (component) in the data (Reynolds 2009). Depending on the restrictions put on the covariance of each Gaussian component, the number of free parameters can go from $N_{\rm G-C}(N_{\rm dim}+1)$ (spherical covariance) to $N_{\rm G-C}(3N_{\rm dim}+1)$ (full covariance). Since the number of components and covariance restriction is given by the user and not *a posteriori* by the algorithm, me 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 values (Schwarz 1978) for a grid of Gaussian components ($N_{\rm G-C} < 20$) and covariance restriction methods for each variable (Figure **??**). We selected the models which yielded the lowest BIC value and plotted the climatological clusters in Figures **??**-**??**. To do this we used the Python `sklearn.mixture.GMM` module (Pedregosa et al. 2011).

Table 2. Climate clusters according to a lowest BIC-based selection of Gaussian Mixture Models.

Variable	Symbol	Lowest BIC No. of Clusters	Lowest BIC Cov. Method	Preferred Clusters
Precipitation	P	11	Full	2,3,7,11
Rainy Days	D	6	Full	1,4,6
Relative Humidity	H	2	Full	2
Sunshine	S	5	Tied	1,4,5

Table 2 shows the selected number of clusters and covariance method for each variable along with our preferred clusters that correlate with dry, sunny climate. 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.

3. Quality Index

We wish to identify all stations for whom all four criteria 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 satisfing 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. 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 for a given station, the total average plus one standard deviation for precipitation, rainy days, and relative humidity correspond to less than 4.3 mm/d, 17 d/mo, and 81% respectively, and for sunshine, to more than 4 h/d. 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 **??**), we propose a quality measure based on a probabilistic analysis of the data.

3.1. Probabilistic Analysis

We base our analysis on the probability for our hypothesis that a station which appears in our shortlist meets all four criteria, given that it is located at a given elevation. We can write down this probability 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 P, D, H, S events "The station belongs to a preferred cluster for variable X", i.e. $C = P \cap R \cap H \cap S$.
- 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 $M_P \cap \neg M_R \cap M_H \cap M_S$
- h is the event that a given station is located at an elevation h.

Thus, the probability that a station $j = (0, 1, ..., N_{\text{stations}})$ 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. whether a station that did not measure a given variable in reality belongs to one of our preferred climatological clusters for that variable. Thus we will use Bayes' theorem to help estimate this probability using simpler conditional probabilities. First we control for the inhomogeneous distribution of elevations $P(h)$,

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

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

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

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,

$$P(C \cap A \cap T_i \mid h) = \frac{P(h \mid C \cap A \cap T_i)}{P(h)} P(C \mid A \cap T_i) P(A \mid T_i) P(T_i) \ . \tag{3}$$

Here $P(h \mid C \cap A \cap T_i$ 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. 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. We can rewrite this probability as,

$$P(C \mid A \cap T_i) = P(P \cap D \cap H \cap S \mid A \cap T_i) \ , \tag{4}$$

$$= \frac{P(P \cap D \cap H \cap S \cap A \cap T_i)}{P(A \cap T_i)} \ . \tag{5}$$

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 rainy days, i.e.,

$$T_1 = M_P \cap M_D \cap \neg M_H \cap \neg M_S \ . \tag{6}$$

If this station of type T_1 is in our shortlist (A), the P and D criteria are met. Thus,

$$P(P \cap D \cap H \cap S \cap A \cap T_1) = P(H \cap S \mid A \cap T_1) \ . \tag{7}$$

Substituting this into Eq. (5),

$$P(C \mid A \cap T_1) = \frac{P(H \cap S \mid A \cap T_1)}{P(A \cap T_1)} = P(H \cap S \mid A \cap T_1) \tag{8}$$

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

$$P(C \mid A \cap T_i) \simeq P(H \cap S) \ . \tag{9}$$

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.

Since the probabilities for stations in our shortlist vary by orders of magnitude, for clarity's sake we decided to use a logarithmic quality index for a station in our shortlist. For $j = (0, 1, ..., N_{\rm stations})$,

$$Q_j = 9 \ \frac{\log(P_{\rm j}/P_{\rm min})}{\log(P_{\rm max}/P_{\rm min})} + 1 \ . \tag{10}$$

Here $P_{\rm min}, P_{\rm max}$ 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 $P_j = P_{\rm max}$, $Q_j = 10$, and if $P_j = P_{\rm min}$, $Q_j = 1$.

4. Results

We selected clusters of climatologically similar stations across Colombia using unsupervised learning methods for each measured variable. From these clusters, 83 weather stations came up as potential candidates for identifying high-mountain regions (at an elevation greater that 2000 masl) with unusually dry, sunny weather (Figure **??**).

Given that many of our candidate stations were lacking in relative humidity and sunshine data, we extrapolated for them using data from nearby stations (at less than 4 km away and at an elevation difference of less than 100 m). We made sure that in general, for stations less than 5 km away, the difference in relative humidity and sunshine data is less than 15%. The extrapolated data suggested that 4 of the stations in the shortlist did not satisfy the relative humidity and sunshine criteria. The remaining 73 stations are plotted in Figure **??**.

All stations outside of the region located within the latitude, longitude range $[(4.3, 6.2), (-72.4, -74.5)]$ seem to be outliers, and while they could indicate unusually dry, sunny regions, there are not enough stations nearby to say anything conclusively. Furthermore, those stations only measured one or two variables, which leads us to reject them from our shortlist.

The final 70 stations located nearby the Cundinamarca-Boyacá region of Colombia are shown in more detail in Figure **??**. This Figure shows two tentative locations in the northern region (Boyacá department) appear, 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 using a lowest-BIC spherical covariance Gaussian Mixture Model on their coordinates (Figure **??**).

The stations are thus classified, and the colored points in Figure **??** show 6 candidate regions. However, in order to see if the prediction from the Gaussian Mixture Models is indicative of regions with similar (dry, sunny) weather, we plotted the predicted variance regions for each cluster along with other stations in our sample (rejects) in Figure **??**. From this figure, we can reject regions where the number of rejected stations within the predicted variance regions is similar to or higher than the number of stations in our shortlist for a specified cluster.

4.1. La Sabana

The region located to the west of Bogotá ()is a hilly region at an elevation above 2600 masl, with a large variety of microclimates, which explains the significant amount of stations from this region in our shortlist. However, Figure **??** shows that within this predicted variance region there are more stations that do not satisfy our criteria than stations that do. Furthermore, stations that meet and do not meet our criteria are homogeneously distributed. 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 for our purposes.

4.2. Valle de Ubaté

This region, located to the NNE of Bogotá s 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 value higher than 6. This indicates a potential candidate region, although more sunshine and humidity data are needed. In our next paper we will study the amount of atmospheric water vapor near this region.

4.3. Villa de Leyva

The region near the town of Villa de Leyva is known for its dry weather, and the stations located within this region seem to support this belief. The ratio of stations not in our shortlist vs. the stations in this predicted variance region is not very low (0.4). However, Figure **??** 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 station is very suggestive 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. This will be covered in our next paper.

4.4. Tunja Canton

The variance predicted 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 regarding this region. However, the high-elevation location of some of these stations is tantalizing, which is why we will keep this as a region of interest for our next paper.

4.5. Valle del Sol

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 value higher than 6, and one of the stations has a very high-Q 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.

4.6. Pisba National Park

This is a *páramo* region, characterized by high-mountain tundra weather. The ratio of stations not in our shortlist to station 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 is needed, but this is not one a strong candidate region.

One, to the NWW of Bogotá, another to the N of Bogotá, in the region known as Valle de Ubaté,

4000 masl might be interesting as well, but there are not enough stations.

References

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 & Duchesnay E 2011 *Journal of Machine Learning Research* **12**, 2825–2830.

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

Schwarz G 1978 *Ann. Statist.* **6**(2), 461–464.
URL: *http://dx.doi.org/10.1214/aos/1176344136*

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.
URL: *http://doi.acm.org/10.1145/1553374.1553494*