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

Germán Chaparro Molano1, Oscar Leonardo Ramírez Suárez1, Oscar Alberto Restrepo Gaitán1,2, and Alexander Marcial Martín Mercado2

1 Grupo de Simulación, Análisis y Modelado, Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia
gchaparrom@ecci.edu.co
2 Radio Astronomy Instrumentation Group, Universidad de Chile, Santiago de Chile, Chile; orestrepog@ecci.edu.co
3 Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá, Colombia
4 Grupo de Simulación del Sistema Climático Terrestre, Universidad Nacional de Colombia, Bogotá, Colombia
5 Departamento de Ciencias Básicas, Universidad ECCI, Bogotá, Colombia

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 six 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 (~10 mm) at an accuracy of ~20 km. Our methods can be applied to similar data sets taken in other countries as a first step toward astronomical site evaluation.

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

Online material: color figures

1. Introduction

The development of astronomical instrumentation technology in the 0.1–2 THz range has been rapidly growing in recent years. Another recent surge in interest in technology at this frequency range has appeared recently due to the saturation of the “classical radio window” in telecommunications (Hosakao 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 altitude regions in the northern Andes could be suitable candidates for a world-class millimeter-wave observatory.

Astronomical observations in the millimeter and sub-millimeter wavelength range require that atmospheric effects affecting absorption at these wavelengths are kept to a minimum (Chamberlin & Grossman 2012; Cortés et al. 2016; Radford & Peterson 2016). The main factor contributing to the atmospheric opacity is water vapor, which very efficiently absorbs light in the THz range (Pardo et al. 2001; 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 (pwv) profile using remote sensing techniques such as microwave radiometry (Peter & Kämpfer 1992; Paine et al. 2000; Battistelli et al. 2012), satellite measurements (Aumann et al.
et al. 2001; Luini & Riva 2016, or indirectly via models of in situ climatological measurements (Lew & Uscka-Kowalkowska 2016) or GPS-delay studies (Bevis et al. 1992; 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 (Clough et al. 1989; Liebe 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).

We estimated the expected zenith opacity above the Colombian Andes (Figure 1) at cm–mm wavelengths considering columns of 2, 10, and 20 mm of precipitable vapor and typical weather conditions. The elevation for the simulated sites in Figure 1 were chosen according to existing millimeter telescope sites (Pico Veleta, Plateau de Bure) (Dutrey 2000). Thus, considering a best-case scenario (below 2 mm pwv), the potential observing frequency bands would be 70–115 GHz, 120–170 GHz, and 206–223 GHz. In a 10–15 mm pwv scenario, the observing bands would be 70–115 GHz and 120–130 GHz. The opacity models were calculated using the CASA (McMullin et al. 2007) asap.opacity.model class, and deriving the pwv value from the 22 GHz K-band zenith opacity (Deuber et al. 2005; Hiriart & Salas 2007).

Although radiometer measurements are desirable for directly measuring the atmospheric water vapor column, their development and/or deployment can be complex and expensive (Pazmány 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 $N_{\text{dim}} = 12$ to $N_{\text{dim}} \leq 3$ 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 (~20 km) water vapor satellite maps (Suen 2016) and find a possibly significant geographical correlation. In an upcoming paper we will analyze 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 data set 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 Data Set

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 2). The variables that we considered relevant to our work were Elevation (in meters above sea level, or masl), Precipitation...
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 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., \( N_{\text{dim}} = 12 \)). This involves

![Figure 2](https://github.com/saint-germain/ideam)

**Figure 2.** Distribution of weather stations in the data set. (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. 2001). (b) Altitude histogram for all stations in the data set. (A color version of this figure is available in the online journal.)

| 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             |

**Table 1** Types of Stations Classified According to Their Reported Variables

*Note.* Y = Measured, N = not Measured.
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 \( N_{\text{dim}} = 12 \) to \( N_{\text{PCA}} \leq 3 \), 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, \( N_{\text{PCA}}(95\%) = 3 \) and for relative humidity \( N_{\text{PCA}}(95\%) = 2 \). To do this we used the Python sklearn.decomposition.PCA module (Pedregosa et al. 2011). To illustrate, Figure 3 shows the relative humidity data projected across two 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 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 \( N_{G-C} \) of Gaussian distributions, each of which represents a cluster in the data (Reynolds 2015). Depending on the restrictions put on the covariance of each Gaussian distribution across the data space, the number of free parameters can go from \( N_{G-C}(N_{\text{dim}} + 1) \) (spherical covariance) to \( N_{G-C}(3N_{\text{dim}} + 1) \) (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 BIC (Schwarz 1978) for a grid of Gaussian clusters \( (N_{G-C} < 20) \) and covariance restriction methods for each variable (Figure 4). 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 5.

From the climatological clusters in Figures 6–14 we selected clusters that indicate a clear-sky, potentially dry 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 (low humidity, precipitation, rain days, and high sunshine duration) for identifying a region as having a clear-sky, potentially dry climate can be met. However, since only 6% of the stations measured all variables (type \( T_8 \) in Table 1), it is not straightforward to accept or reject stations based on their belonging to one of our selected clusters. In other words, a station that meets less than all four criteria can still be considered in our analysis, but it will not be given the same level of importance as a station that meets all four criteria. This requires setting up a method to determine the quality of a site depending on its location and station type. Thus we selected stations that meet criteria for (and only for) the variables they measured. Thus, if a station measured precipitation and humidity (Type \( T_3 \)), it had to belong to a precipitation and humidity cluster listed in Table 2 or else it was rejected. 665 stations across Colombia showed such behavior.

In order to estimate how much consideration we should give to a shortlisted station with an incomplete measurement set, we propose the use of a quality index based on a probabilistic analysis of the entire data set, accounting for the non-uniform joint distribution of station types (Table 1) and elevations (Figure 2(b)). Neglecting to include all stations in the calculation of this joint probability leads to very strong selection biases, assigning unfairly high quality indices to substandard stations. Therefore, even though we will later limit our selection to high-elevation sites in order to achieve lower pwv regions, we considered all the...
information provided by the full data set in our probabilistic analysis.

4.1. Bayesian Probabilistic Analysis

We quantify the quality of a given site with 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 and is of a given type. We can write down this probability \( P_j \) in terms of the following events:

1. \( C \) 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.
   (a) \( C \) 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).
2. \( A \) is the event that a given station appears in our shortlist.
3. \( T_i \) is the event that a given station is of type \( T_i \) (See Table 1).

Thus, the probability that a station \( j = \{0, 1, \ldots, N_{\text{shortlist}}\} \) 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

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

(a) For example, if \( M_X \) is the event “the station measured variable \( X \),” then the event \( T_i \) is equivalent to \( M_R \cap -M_D \cap M_H \cap M_S \).
The joint probability of all events $C \cap A \cap T_i \cap h$ can be rewritten as

\[
P_j = \frac{P(h \mid C \cap A \cap T_i)}{P(h)} P(C \cap A \cap T_i).
\]  

(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,

\[
P_j = \frac{P(h \mid C \cap A \cap T_i)}{P(h)} P(C \mid A \cap T_i) P(A \mid T_i) P(T_i).
\]  

(3)

where

1. $P(h \mid C \cap A \cap T_i)$ is the distribution of elevations for a station type ($T_i$) on our shortlist ($C \cap A$),
2. $P(A \mid T_i)$ is the probability that a station of type $T_i$ is in our shortlist,
3. $P(T_i)$ is the probability that a station type is $T_i$, and
4. $P(C \mid A \cap T_i)$ 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.\footnote{Distributions of elevation-related probabilities were approximated discretely using a bin size estimated from the mean elevation difference between stations when sorted by elevation.}

\[
P(C \mid A \cap T_i) = \frac{P(C \cap A \cap T_i)}{P(A \cap T_i)}.
\]  

(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.,

\[
T_1 = M_R \cap M_D \cap -M_H \cap -M_S.
\]  

(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,

\[
P(C \cap A \cap T_1) = P(H \cap S \mid A \cap T_1).
\]  

(6)
Substituting this into Equation (4),

\[ P(C \mid A \cap T_i) = \frac{P(H \cap S \mid A \cap T_i)}{P(A \cap T_i)} = P(H \cap S \mid A \cap T_i). \]  

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,

\[ P(C \mid A \cap T_i) \approx P(H \cap S). \]  

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

\[ Q_j = 9 \log \left( \frac{P_j}{P_{\min}} \right) + 1. \]  

Here, \( P_{\min}, P_{\max} \) are the lowest and highest values for the probabilities for the stations in the first shortlist 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

As mentioned above, as we go higher in elevation, the water vapor column decreases. For this reason we included an additional criterion for elevation limiting our list to 119 stations located above 2000 masl. This list was further reduced by
Figure 7. Climate clusters (1–3 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 altitude of weather stations in each precipitation cluster. (A color version of this figure is available in the online journal.)
Figure 8. Climate clusters (4–6 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 altitude of weather stations in each precipitation cluster. Continuation of Figure 7.

(A color version of this figure is available in the online journal.)
Figure 9. Climate clusters (7–9 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 altitude of weather stations in each precipitation cluster. Continuation of Figure 8.

(A color version of this figure is available in the online journal.)
requiring that for any given station, the total average plus 1σ for precipitation, rain days, and relative humidity is respectively less than 4.3 mm day$^{-1}$, 17 days month$^{-1}$, and 81%, and for sunshine duration, more than 4 hr day$^{-1}$. This left us with a shortlist of 83 weather stations.

Many of the stations in our data set are lacking in relative humidity and sunshine duration data. In order to see if we could extrapolate using data from nearby stations, we estimated the difference between relative humidity and sunshine duration data for all stations in the entire data set. For stations at a relative elevation of less than 100 m and located at less than 4 km away from each other, this difference amounts to less than 15%. Thus, we extrapolated data for some of the 83 candidate stations in our shortlist, and were able to reject 4 stations which did not satisfy the relative humidity and sunshine duration criteria. The remaining 79 stations are plotted in Figure 15. (A color version of this figure is available in the online journal.)

We rejected 9 additional stations located outside of the region located within the latitude, longitude range $[4.3, 6.2]$, $[-72.4, -74.5]$ (Figure 15(a)) due to geographical sparsity. Even though they could be indicative of clear-sky, potentially dry regions, there are not enough stations nearby to make any strong conclusions regarding those regions, as shown in Figure 16. Adding to this, in all 9 cases the only variable they measured was precipitation ($Q_j = 2.2$), which led 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 15(b). This figure shows two candidate locations in the northern
Figure 11. Climate clusters (1–3 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 altitude of weather stations in each rain days cluster. (A color version of this figure is available in the online journal.)
Figure 12. Climate clusters (4–6 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 altitude of weather stations in each rain days cluster. Continuation of Figure 11. (A color version of this figure is available in the online journal.)
Figure 13. Climate clusters (1–3 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 altitude of weather stations in each sunshine duration cluster. (A color version of this figure is available in the online journal.)
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 17).

### 5. Results

We selected clusters of climatologically similar stations across Colombia using an unsupervised learning low-dimensional embedding algorithm for each measured variable. From

![Figure 14. Climate clusters (4–5 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 altitude of weather stations in each sunshine duration cluster.](image)

![Table 2](image)

| 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           |

Note: The selected clusters column refers to the climate clusters shown in Figures 6–11.
these clusters, 70 weather stations came up as potential candidates for identifying high-mountain regions (at an elevation greater than 2000 masl) with clear-sky, potentially dry weather. The driest months of the year for all the candidate regions of interest are December–February and June–August (Figures 6–14). It should be noted that humidity in the Colombian Andes can change rapidly during the course of the day (Pinzón et al. 2015), and daytime relative humidity measurements <50% are often reported in high-mountain plateau stations.

5.1. Candidate Regions of Interest

Figure 17 shows six candidate regions of interest, where weather stations indicating clear-sky, potentially dry 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 (red circles), and plotted along rejected stations in our original data set (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.

(A color version of this figure is available in the online journal.)

![Diagram](image)

**Figure 17.** Geographical clustering results for a lowest-BIC spherical covariance Gaussian Mixture Model of the stations located inside the red circle in Figure 15(b), indicating the predicted variance regions for each cluster (red circles), and plotted along rejected stations in our original data set (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.

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 17. 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°77N, 74°18W) and covers an radius of 15.3 km (Figure 17, magenta 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 17 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.

### 5.1.2. Valle de Ubaté

This cluster is located at (5°30N, 73°80W) and covers a radius of 11.4 km (Figure 17, cyan 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°61N, 73°55W) and covers a radius of 7.11 km (Figure 17, 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 versus the stations in this predicted variance region is not very low (0.4). However, Figure 17 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°59N, 73°24W) and covers a radius of 8.29 km (Figure 17, purple 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°72N, 72°96W) and covers a radius of 8.06 km (Figure 17, 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°04N, 72°71W) and covers a radius of 10.52 km (Figure 17, teal 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. Even though raw MODIS data has a potentially higher spatial resolution, the resolution of the processed maps is not fine enough to pinpoint a site location at a precision higher than $\sim$20 km. The area with the lowest mean precipitable water vapor that MODIS is able to measure ($\sim$10–15 mm) in the Suen et al. (2014) map for Colombia spans the Tunja, Valle del Sol, Pisba, and part of the Valle de Ubaté regions. However, it also spans nearby regions with rejected stations, which means that low-resolution precipitable water vapor maps can indicate a wide, first order candidate region, but fail to account for highly local weather conditions, which are accounted for in our method. For this reason, in our next paper we will correlate the results presented here with existing GPS-delay water vapor measurements from the Colombian Global Navigation Satellite System GeoRED network (https://geored.sgc.gov.co) and seasonal, higher resolution MODIS data.

6. Conclusions

In this paper we find plausible locations for a high-mountain cm- to mm-wave astronomical observatory in the northern Andes of Colombia, by analyzing 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 data set 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 six 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 (December–February and June–August) 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. 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). The La Sabana region is also not a good candidate due to a high rejection rate of stations within this region.

We summarize the information about two regions of interest (Valle de Ubaté and Valle del Sol) in Table 3 with the highest Bayesian probabilistic indices, reflecting the quality of the data for stations in those regions. Most of the stations in this table
are located at elevations 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 ∼10–15 mm in ∼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 could warrant further measurements using GPS stations, radio sondes and radiometers. The same could be said for the Pisba and Tunja regions, but the low quality of the climate data (indicated by low Bayesian probabilistic indices) indicates that more data is needed. For this reason, we will correlate the results for these four regions with GPS-delay water vapor measurements in our following paper.

There is an additional issue with the Valle del Sol region. The high industrial activity in this region fills the air with particulate matter (Jimenez Pizarro 2004; Jimenez Pizarro et al. 2011), which would nullify the suitability of this region. Besides obtaining seasonal, higher-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 climatological data sets in other countries. Given a similar spatial and temporal coverage, our methods can provide a better picture of local climate than global satellite maps, and can help narrow down plausible locations for direct atmospheric opacity measurement campaigns. This can be a low-cost, preliminary step in radio astronomy site testing in developing countries.

This work was made possible by the Vicerrectoría de Investigación 01-2015 research grant from Universidad ECCI. We would also thank the Instituto de Hidrología, Meteorología

### Table 3

| Region of Interest | Station Code | Municipality | Qj | rc (km) | \(\bar{R}\) (mm month\(^{-1}\)) | \(\bar{D}\) (days month\(^{-1}\)) | \(\bar{R}\) (%) | \(\bar{S}\) (hr\(^{-1}\) day\(^{-1}\)) | h (masl) |
|--------------------|--------------|--------------|-----|--------|-------------------------------|----------------|-----------|----------------|---------|
| Valle de Ubaté     | 21201410     | Nemocon      | 2.2 | 28     | 66                            | ND             | ND        | ND             | 2600    |
|                    | 240111060    | Susa         | 4.6 | 18     | 83                            | 12             | ND        | ND             | 2600    |
|                    | 24010140     | Cucunuba     | 4.7 | 6      | 57                            | 9              | ND        | 2620           |         |
|                    | 24015210     | Sutatausa    | 4.8 | 6      | 57                            | 10             | ND        | 2700           |         |
|                    | 24010170     | Guacheta     | 4.9 | 16     | 73                            | 10             | ND        | 2690           |         |
|                    | 24010070     | Lenguazaque  | 5.2 | 10     | 63                            | 12             | ND        | 2650           |         |
|                    | 24015130     | Simijaca     | 5.2 | 23     | 67                            | 11             | ND        | 2572           |         |
|                    | 21201620     | Suesca       | 5.6 | 20     | 59                            | 10             | ND        | 2575           |         |
|                    | 24010610     | Carmen de Carupa | 6.1 | 12   | 54                            | 12             | ND        | 2970           |         |
|                    | 24011080     | Cucunuba     | 6.1 | 7      | 51                            | 5              | ND        | 2562           |         |
|                    | 24010280     | Lenguazaque  | 6.1 | 9      | 57                            | 10             | ND        | 2585           |         |
|                    | 24010440     | Susa         | 6.1 | 11     | 60                            | 8              | ND        | 3130           |         |
|                    | 24011090     | Ubate        | 6.1 | 4      | 62                            | 13             | ND        | 2555           |         |
|                    | 24015110     | Ubaté        | 6.1 | 5      | 62                            | 12             | 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        | 2500           |         |
|                    | 24035140     | Sogamoso     | 5.0 | 4      | 59                            | 10             | ND        | 2500           |         |
|                    | 24030940     | Sogamoso     | 5.0 | 7      | 60                            | 12             | ND        | 2500           |         |
|                    | 24034011     | Tibasosa     | 5.0 | 5      | 62                            | 7              | ND        | 2500           |         |
|                    | 24030410     | Iza          | 5.6 | 12     | 52                            | 9              | ND        | 2470           |         |
|                    | 24030540     | Firavitoba   | 6.1 | 6      | 57                            | 9              | ND        | 2486           |         |
|                    | 24030190     | Mongui       | 6.1 | 12     | 65                            | 9              | ND        | 2970           |         |
|                    | 24030760     | Sogamoso     | 6.1 | 11     | 65                            | 10             | ND        | 3225           |         |
|                    | 24035340     | Sogamoso     | 6.7 | 5      | 61                            | 12             | 75        | 5              | 2500    |
|                    | 24035150     | Nobsa        | 10.0| 10     | 68                            | 13             | 74        | 4              | 2530    |

Note. Detailed table (including all clusters) available at https://github.com/saint-germain/ideam.
y Estudios Ambientales (IDEAM) for providing us with the data set, and the anonymous referee for helping us convey our ideas more clearly.

**ORCID iDs**

Germán Chaparro Molano @ https://orcid.org/0000-0003-0919-1512

**References**

Archibald, E. N., Jenness, T., Holland, W. S., et al. 2002, MNRAS, 336, 1
Aumann, H. H., Chahine, M. T., Gautier, C., et al. 2003, ITGRS, 41, 253
Bevis, M., Businger, S., Herrington, T. A., et al. 1992, JGRD, 97, 15
Battistelli, E. S., Amico, G., Ba, A., et al. 2012, MNRAS, 423, 1293
Bustos, R., Rubio, M., Otrola, A., & Nagar, N. 2014, PASP, 126, 1126
Caumont, O., Cimini, D., Löhner, U., et al. 2016, QJRMS, 142, 2692
Chamberlin, R. A., & Grossman, E. N. 2012, JGRD, 117, D13111
Cimini, D., Westwater, E. R., Gasiewski, A. J., et al. 2007, ITGRS, 45, 2169
Clough, S. A., Kneizys, F. X., & Davies, R. W. 1989, AtmRe, 23, 229
Clough, S. A., Kneizys, F. X., & Davies, R. W. 1996, AtmRe, 24, 1
Clough, S. A., Kneizys, F. X., & Davies, R. W. 2002, AtmRe, 30, 1
Cortés, F., Reeves, R., & Bustos, R. 2016, RaSc, 51, 1166
Denny, S., Suen, J., & Lubin, P. 2013, NewA, 25, 114
Deuber, B., Morland, J., Martin, L., & Kämpfer, N. 2005, RaSc, 40, 1
Dutrey, A. (ed.) 2000, in Millimeter Interferometry Summer School 2”, ed. A. Dutrey (France: St. Martin d’Hères), http://iram.fr/Iramfri/IS/IS2002/html_2/book.html
He, J., Zhang, S., & Wang, Z. 2012, RaSc, 47, RS2009
Hiriart, D., & Salas, L. 2007, RMxAA, 43, 225
Hosok, I., Sekine, N., Patrashin, M., et al. 2007, Proc. IEEE, 95, 1611
Jimenez Pizarro, R. 2004, PhD thesis, École Polytechnique Fédérale de Lausanne
Jimenez Pizarro, R., Arango, C. D., & Peña, J. A. 2011, AGUFM, A51C-08
Jones, A., Walker, K. A., Jin, J. J., et al. 2012, ACP, 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., Bihler, S., & Knzi, K. 2002, JQSRT, 74, 545
Lew, B., & Uscka-Kowalkowska, J. 2016, MNRAS, 455, 2901
Liebe, H. J. 1989, IJIMW, 10, 631
Liljegren, J. C., Clothiaux, E. E., Mace, G. G., Kato, S., & Dong, X. 2001, JGRD, 106, 14485
Luini, L., & Riva, C. G. 2016, ITAP, 64, 2487
McMullin, J. P. et al. (ed.) 2007, in ASP Conf. Ser. 376, Astronomical Data Analysis Software and Systems XVI (San Francisco, CA: ASP), 127
Nelli, A. E., Coster, A. J., Solheim, F. S., et al. 2001, JATOM, 18, 830
Paine, S., Blundell, R., Papa, D. C., Barrett, J. W., & Radford, S. J. 2000, PASP, 112, 108
Pardo, J., Serabyn, E., & Cernicharo, J. 2001, JQSRT, 68, 419
Pardo, J. R., Cernicharo, J., & Serabyn, E. 2001, ITGRS, 45, 2208
Pazmany, A. L. 2007, ITGRS, 45, 2202
Pedregosa, F., Varoquaux, G., Gramfort, A., 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, JGRD, 97, 18173
Pickett, H. M., Poynter, R. L., Cohen, E. A., et al. 1998, JQSRT, 60, 883
Pinzón, G., González, D., & Hernández, J. 2015, PASP, 127, 523
Radford, S. J. E., & Peterson, J. B. 2016, PASP, 128, 075001
Reynolds, D. 2015, Gaussian Mixture Models (Boston, MA: Springer)
Rosenkranz, P. W. 1998, RaSc, 33, 919
Seidel, D. J., Fu, Q., Randel, W. J., & Reichler, T. J. 2008, NatGe, 1, 21
Shaw, B., & Jebara, T. 2009, in Proc. 26th Annual Int. Conf. on Machine Learning (New York: ICML ’09 ACM), 937 doi:10.1145/1553374.1553494
Slocum, D. M., Slingerland, E. J., Giles, R. H., & Goyette, T. M. 2013, JQSRT, 127, 49
Smith, G. J., Naylor, D. A., & Feldman, P. A. 2001, IJIMW, 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, ITTST, 4, 86
Turner, D. D., Cadeddu, M. P., Lohnert, U., Crewell, S., & Vogelmann, A. M. 2009, ITGRS, 47, 3526
Wang, J., Zhang, L., Dai, A., Van Hove, T., & Van Baelen, J. 2007, JGRD, 112, D11107
Wentz, F. J., & Meissner, T. 2016, RaSc, 51, 381
Withayachumnankul, W., Fischer, B. M., & Abbott, D. 2008, RSPSA, 464, 2435