The Use of Euclidean Geometric Distance on RGB Color Space for the Classification of Sky and Cloud Patterns

SYLVIO LUIZ MANTELLI NETO

Earth System Sciences Center (CCST-INE), National Institute for Space Research, São José dos Campos, São Paulo, and Knowledge and Engineering Department (EGC), and Image Processing and Graphics Computing Lab (LAPIX), and Solar Energy Lab (LABSOLAR), Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil

ALDO VON WANGENHEIM

Image Processing and Graphics Computing Lab (LAPIX), Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil

ENIO BUENO PEREIRA

Earth System Sciences Center (CCST-INE), National Institute for Space Research, São José dos Campos, São Paulo, Brazil

EROS COMUNELLO

University of Itajai Valley (UNIVALI), São José, Santa Catarina, Brazil

(Manuscript received 1 July 2009, in final form 17 November 2009)

ABSTRACT

The current work describes the use of multidimensional Euclidean geometric distance (EGD) and Bayesian methods to characterize and classify the sky and cloud patterns present in image pixels. From specific images and using visualization tools, it was noticed that sky and cloud patterns occupy a typical *locus* on the red–green–blue (RGB) color space. These two patterns were linearly distributed parallel to the RGB cube's main diagonal at distinct distances. A characterization of the cloud and sky patterns EGD was done by supervision to eliminate errors due to outlier patterns in the analysis. The exploratory data analysis of EGD for sky and cloud patterns showed a Gaussian distribution, allowing generalizations based on the central limit theorem. An intensity scale of brightness is proposed from the Euclidean geometric projection (EGP) on the RGB cube's main diagonal. An EGD-based classification method was adapted to be properly compared with existing ones found in related literature, because they restrict the examined color-space domain. Elimination of this limitation was considered a sufficient criterion for a classification system that has resource restrictions. The EGD-adapted results showed a correlation of 97.9% for clouds and 98.4% for sky when compared to established classification methods. It was also observed that EGD was able to classify cloud and sky patterns invariant to their brightness attributes and with reduced variability because of the sun zenith angle changes. In addition, it was observed that Mie scattering could be noticed and eliminated (together with the reflector's dust) as an outlier during the analysis. Although Mie scattering could be classified with additional analysis, this is left as a suggestion for future work.

1. Introduction

Automatic cloud evaluation from the surface is an important issue in meteorology to reduce subjective aspects and operational costs of synoptic observers (SO). Several research groups are demanding new techniques for automatic cloud and sky detection to replace SO using automatic cameras (World Climate Research Program 2007). Substituting the SO evaluation with the automatic system defined by the World Meteorological Organization (2008) as a synoptic observation system (SOS) is not a trivial task. It involves aspects of human perception, atmospheric sciences, mathematics, computer artificial intelligence, etc., in the design of an ''artifact'' called an intelligent agent (IA; Russell and Norvig 2003, chapter 2). The existing image analysis artifacts used as SOS do not match the qualitative performance of SO, and to find

Corresponding author address: Sylvio Mantelli Neto, UFSC-EMC-LABSOLAR, Campus Trindade, Florianopolis SC Brazil, 88040-900.
E-mail: sylvio@lepten.ufsc.br

DOI: 10.1175/2010JTECHA1353.1

better solutions, improvements to the classification techniques must be made.

SO observations are normalized by the World Meteorological Organization (2008, chapter 15). These observations usually describe cloud type and amount, but the evaluation is highly subjective. With regards to cloud amount, for instance, inconsistencies exist when the same sky is evaluated by distinct operators (Hoyt 1978). The packing effect is another inconsistency, which is caused by an overestimation of the amount of clouds near the horizons (Holle and Mackay 1975). The natural human lack of consistency, due to the operator's fatigue and the effects of shifts of observation teams, stimulate the use of automatic systems. However, most camera-based SOS methods are still compared to SOs for their validation on the qualitative analysis of sky conditions (Souza-Echer et al. 2006; Long et al. 2006).

Clouds can be evaluated from satellite images, avoiding some of the previously mentioned problems but introducing others (Rossow 1982)—such as the mentioned pixel geometric distortion, clouds that cannot be detected between layers because of the vertical distribution, seasonal surface variation, etc. Any satellite-based assessment (e.g., energy balance, temperature, radiation, wind, clouds, etc.) must take into account the surface observations' "ground truths" (GTs) to reduce modeling uncertainties. Furthermore, there is a consensus that for cloud evaluation a complementary observation will be the best way to reduce the limitations of both techniques.

SOs are trained to develop cognitive skills for sky pattern identification, but perform poorly when determining the precise amount of clouds. Camera-based SOSs have a better performance when determining cloud amounts than a surface observer, but perform worse for pattern identification. A Camera-based SOS also relies on methodologies like simplified dimension thresholding (Souza-Echer et al. 2006) and the reduction of multivariate color spaces (Long et al. 2001). Those were the only methodologies found in the literature that used image analysis. Short wave, long wave, and other sensors used for that purpose are out of the scope of the current work, but they could be investigated as a cross-comparison analysis in the future. The present work will consider only the comparison of the two equivalent methods existing in the literature for surface image analysis.

Souza-Echer et al. (2006) used a flat image with a 62° field of view (FOV) camera on a zenith mount, always avoiding the direct sun light. Only the saturation dimension of hue, saturation, and lightness (HSL, cylindrical coordinates) was used on the characterization of the three patterns: sky, clouds, and a third class obtained by exclusion. The discrimination function to classify sky and clouds is based on three standard deviation level thresholding from the pattern average. Only these three patterns and their amount are produced by this approach. The elimination of the sun from the observation domain restricts the analysis to small brightness patterns only.

Long et al. (2006) employed a different criteria and an experimental setup using two pieces of equipment: the Total Sky Imager (TSI) and the Whole Sky Camera (WSC) . TSI uses an image from a reflected mirror and WSC a direct image of the sky, both with a 160° FOV. A detailed description about the experimental setup and analysis is provided by Long et al.'s (2006) paper and in Long et al. (2001). Although the cameras obtained images in a 24 bits per pixel red–green–blue (RGB) file format, the classification is restricted to 0.6 threshold R–B dimensions ratio (Long et al. 2006). This criterion reduces the domain color analysis from black to magenta only (Gonzales and Woods 2002), discarding any reference or additional data that could be gathered from the green channel, which might help to classify or analyze further information of atmospheric patterns seen from images. The paper also points out the difficulty of identifying small differences in patterns due to atmospheric contents. Brighter-blue pixels representing blue skies in the transition between molecular scatterings to turbidity are difficult to classify with the proposed method, and they probably would be difficult to detect from a reduced dimension thresholding classification (Mantelli Neto 2001, 2005).

The aforementioned methodologies represent some important pioneering techniques aiming at the replacement of sky-state observers in meteorological stations. However, both methodologies do not use all the possibilities available for image analysis. A 24-bit image allows $2^{24} = 16\ 777\ 216$ different color combinations that can be grouped, analyzed, and combined with a great potential to be explored as a domain. Souza-Echer et al. (2006) use only 8 bits or $2^8 = 256$ lightness possibilities to identify clouds. Long et al. (2001) use 16 bits or $2^{16} = 65\ 536$ colors. Clouds, however, are white with equally likely components of red, blue, and green. Lillesand and Kiefer (1994) define clouds as nonselective, equally scattering all color components. Our work also presents a new point of view on sky pattern analysis based on a Bayesian methodology (Tenenbaum et al. 2006; Chater et al. 2006; Russell and Norvig 2003, chapter 20; National Institute Of Standards 2010, section 1) to improve surface automatic observation of the atmosphere.

The methodology of the present work uses multivariate color space features (Johnson and Wichern 2007) to classify clouds and sky pixels by means of a pattern statistical characterization using the Euclidean geometric distance (EGD). It also intends to propose a scale based on the brightness projection from the Euclidean geometric

projection (EGP) value of pixels on the RGB diagonal cube.

The next sections will show in a Bayesian approach the exploratory data analysis (EDA) of the pixel pattern domain and the mathematical approach to find the geometric position of the target patterns. From that approach, a solution was implemented on a graphical user interface (GUI) and input images were analyzed using the proposed method. A preliminary analysis and a color-space dimension reduction have been made to allow a comparison with other methods.

2. Material and methods

a. Experimental setup and preliminary analysis

Images were collected using a commercially available sky imager (TSI-440, available online at http://www.yesinc.com/products/data/tsi440/index.html) in standard Joint Photographic Experts Group (JPEG) file format at a 352 × 288 resolution with 24-bit colors. Images were obtained not directly from the sky, but by a dome-shaped reflector every 15 min from dawn to dusk, according to TSI program parameters. The reflected image represents a 160° hemispheric angle of view. An adhesive, moving shading band was applied on the reflector surface to avoid damage on the camera by direct exposition to sunlight. A preliminary image processing was performed on images to eliminate spurious and systematic patterns that might interfere in the image analysis (Montgomery 2005). Three patterns are defined as spurious: border effects, horizontal obstructions, and the moving shading band. Borders are not relevant for image analysis because they represent the equipment's self image. Horizontal obstructions are fixed objects like poles, towers, trees, buildings, and geographic features that are present in the image and do not represent any useful information. The equipment's moving shading band was also masked. Obstructions and the horizon have a fixed position and are easily masked on the image. But the shading band is a dynamic feature that moves according to the solar movement and for every image. A different mask file was obtained according to each specific time position. Masking was performed by hand-marking patterns pixel by pixel using the software GT generator tool developed by the research group. The output of the GT tool is a black mask in a bitmap (BMP) file. Mask files are loaded dynamically by the GUI software tool during the analysis phase for every input image.

The images were obtained in the facilities of the Solar Energy Laboratory of the Federal University of Santa Catarina, Florianopolis, Brazil, (LABSOLAR, available online at http://www.lepten.ufsc.br/) located at 27°32′S, 48°31′W. The site is also a Baseline Surface Radiation Network (BSRN, available online at http://www.bsrn.awi.de/) station site. A set of preliminarily images were analyzed on the RGB and HSL color spaces using the Color Inspector 3D visualization tool (available online at http://www.f4.htw-berlin.de/~barthel/ImageJ/ImageJ.htm).

During visual inspection of several collected images, a typical *locus* of cloud and sky patterns can be noticed in the color space. The typical locus related to pattern presence is illustrated in Fig. 1 for a cloudy sky and Fig. 2 for a blue sky. Color inspector 3D showed that cloud pixels are typically gray and white, distributed linearly, and closely parallel to the RGB diagonal cube. It is also possible to notice a luminous gradient-scale distribution along the main diagonal. Sky patterns also showed a linear behavior in the RGB cube, but pixels were located a bit farther from the RGB diagonal. Sky and cloud patterns of the same images showed a nonlinear behavior in the HSL color space. Based on those observations, it was decided to perform mathematical operations in the RGB color space, avoiding the HSL nonlinearity. But some analyses were still performed on HSL because its representation is more closely related to the human perception of colors. The other color spaces were not considered.

b. Cloud and sky pattern characterization with exploratory data analysis

A preliminary statistical analysis was used on EGD to characterize the target patterns identified visually at two different distances from the RGB main diagonal. The selected images represent typical samples of cloud and sky patterns. The same original cloud and sky images seen on Figs. 1 and 2 (and Figs. 4 and 5) were used as reference ground truths (Jiang et al. 2006) for those two patterns (Fernandez-Garcıa et al. 2008). Then an EDA was performed for each pattern to determine their typical EGD from RGB cube diagonal.

To explain the EGD method, a generic pixel in the RGB color space is shown in Fig. 3. the pixel distances can be determined by considering them as vector modules. Projection and distance of pixels from the main diagonal can be easily calculated by

$$\mathrm{PROJ} = |\overline{\mathbf{D}}| \cos(a),$$
$$\mathrm{DIST} = |\overline{\mathbf{D}}| \sin(a).$$

By means of the Al-Kashi theorem (also known as the law of cosines) and the sum of vectors formula we obtain

$$|\overline{\mathbf{C}}| = \sqrt{255^2 + 255^2 + 255^2} = 441.673,$$
$$|\overline{\mathbf{A}}|^2 = |\overline{\mathbf{D}}|^2 + |\overline{\mathrm{X}}|^2 - 2|\overline{\mathbf{D}}||\overline{\mathbf{C}}| \cos(a),$$
$$\overline{\mathbf{C}} = \overline{\mathbf{D}} + \overline{\mathbf{A}},$$

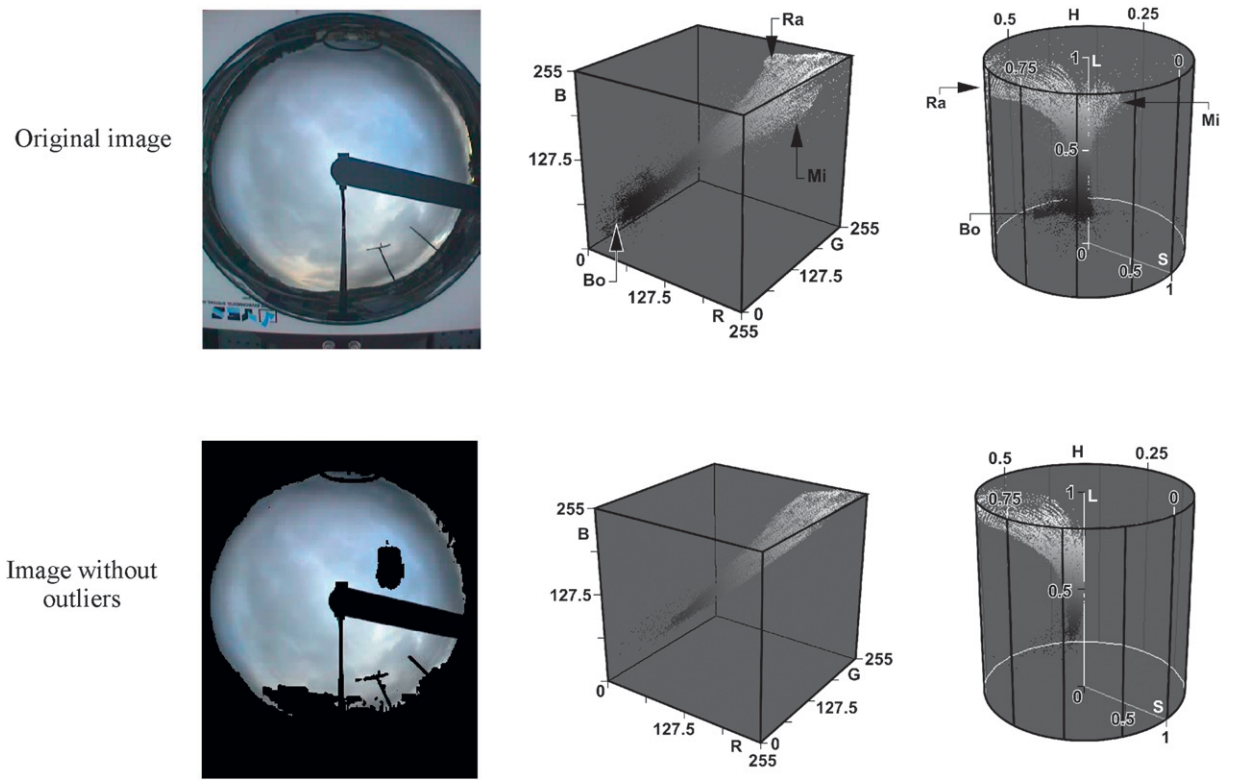

FIG. 1. (a) The typical locus of the cloud patterns observed in images are shown in (b) RGB color-space and (c) HSL color space. The first row is related to the original image and the second row to masked images without outliers. Outliers were masked to black. "Ra," "Mi," and "Bo" labels on the color space indicate, respectively, the typical locus of Rayleigh- and Mie-scattering patterns, the equipment's border, and the shading band. Units are in pixel relative intensity for column (b) and pixel-normalized relative intensity for column (c).

where

- a is the angle between the analyzed vector pixel value and the main diagonal;
- $\overline{\mathbf{A}}$ is the complementary vector from the color pixel to the RGB cube vertex;
- $\overline{\mathbf{C}}$ is the color cube main diagonal vector, with a value of (255 255 255); and
- $\overline{\mathbf{D}}$ is the pixel having (R, G, and B) values.

Replacing $\overline{\mathbf{A}}$ in the Al-Kashi theorem and rearranging the formula leads to

$$\cos(a) = \frac{\left|\overline{\mathbf{C}} - \overline{\mathbf{D}}\right|^2 - \left|\overline{\mathbf{D}}\right|^2 - \left|\overline{\mathbf{C}}\right|^2}{-2\left|\overline{\mathbf{D}}\right|\left|\overline{\mathbf{C}}\right|}.$$

The formulation above demonstrates that only the pixel value is necessary to calculate its distance and projection in the main RGB diagonal. The distance values' text files were generated in a GUI interface from images and loaded into a commercially available statistical analysis tool. After being analyzed, the distances of the two patterns were considered as normally distributed continuous variables, with a typical mean and variance distance from the cube's diagonal. Image characterization for the cloud pattern obtained from RGB coordinates are illustrated in Fig. 4.

To eliminate outlier points, the interquartile distance (IQD) range was used (which is the distance between the upper and lower quartile of data). The range is obtained from the calculated points of the lower 25% (Q1) and the upper 75% (Q3) of data statistics. Points outside the inferior and superior interquartile limits (IQI and IQS, respectively) of the average point were eliminated. A summary of the equations used in outlier elimination is described below:

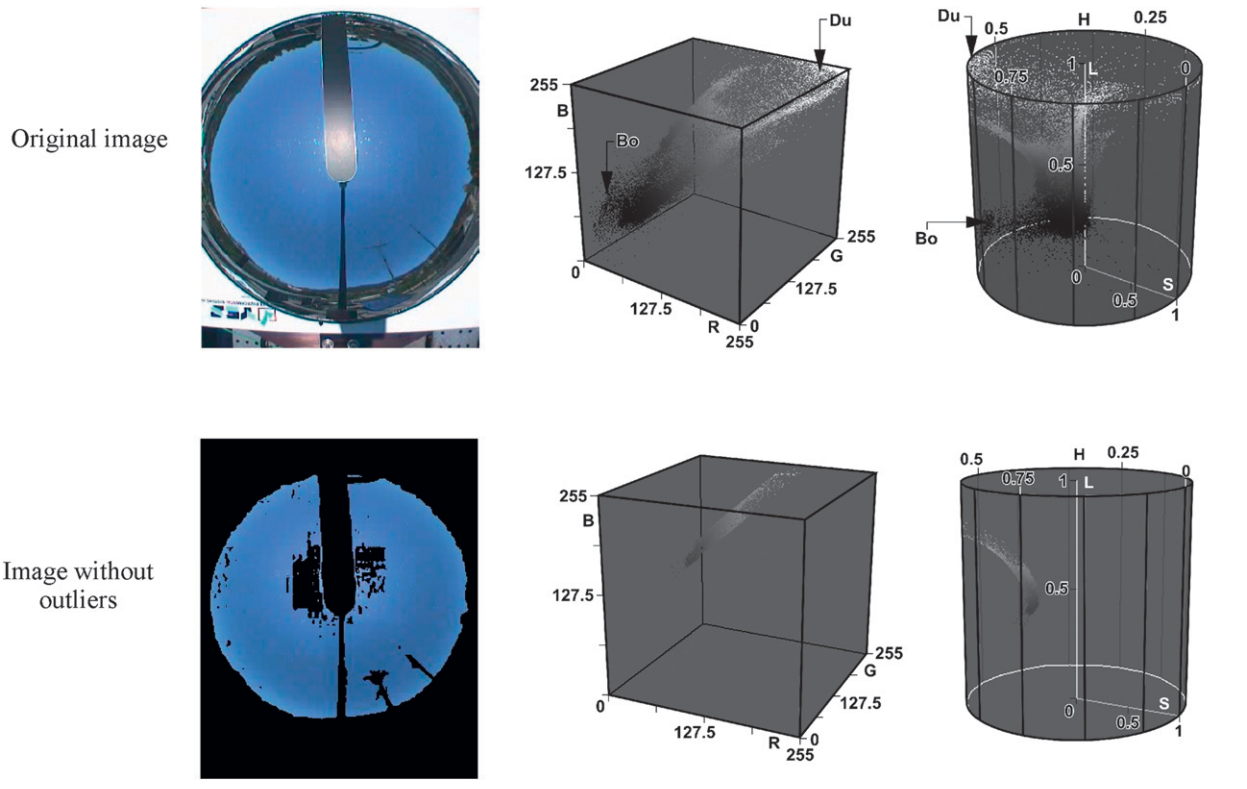

FIG. 2. (a) The typical locus of sky patterns observed in images are shown in (b) RGB color-space column and (c) HSL color space. The first row is related to the original image and the second row to masked images without outliers. Outliers were masked to black. "Du" and "Bo" on the color space indicate, respectively, the typical locus of Dust pattern and the equipment's border and shading band. Units are in pixel relative intensity for column (b) and pixel normalized relative intensity for column (c).

IQD = Q3 – Q1, IQD: Interquartile distance,
IQI = Q1 – 1.5 × IQD, IQI: Inferior interquartile limit, and
IQS = Q3 + 1.5 × IQD, IQS: Superior interquartile limit.

Figure 4 illustrates the statistical analysis that was performed with the pixel values extracted from masked cloud images and the pixel values converted into text files. Outlier extraction is made in the following way by using a statistical analysis

$$\text{IQD} = \text{Q3} - \text{Q1} = 12.75 \quad \text{and}$$
$$\text{IQS} = \text{Q3} + 1.5 \times \text{IQD} = 45.315.$$

Cloud outlier pixels were discarded if distance >45.315. Only the superior interquartile limit value was considered because the distribution of clouds pixels starts in the main diagonal. After excluding the outlier values from cloud images, the remaining pixels from the original image can be visualized on the second row of Fig. 1. A summary for cloud pattern limits is illustrated in Table 1.

For the sky pattern, pixel values from the masked image were extracted and converted into text files to perform the statistical analysis shown in Fig. 5. From analysis of the figure, a statistical outlier extraction was developed using the following conditions:

$$\text{IQD} = \text{IQ3} - \text{IQ1} = 8.26,$$
$$\text{IQS} = \text{IQ3} + 1.5 \times \text{IQD} = 89.240, \quad \text{and}$$
$$\text{IQI} = \text{IQ1} - 1.5 \times \text{IQD} = 56.200.$$

Clear-sky outlier pixels were discarded if they were not in the interval 56.200 ≤ distance ≤ 89.240. After excluding outliers from the image, the remaining pixels can be visualized in the image of Fig. 2 in the second row.

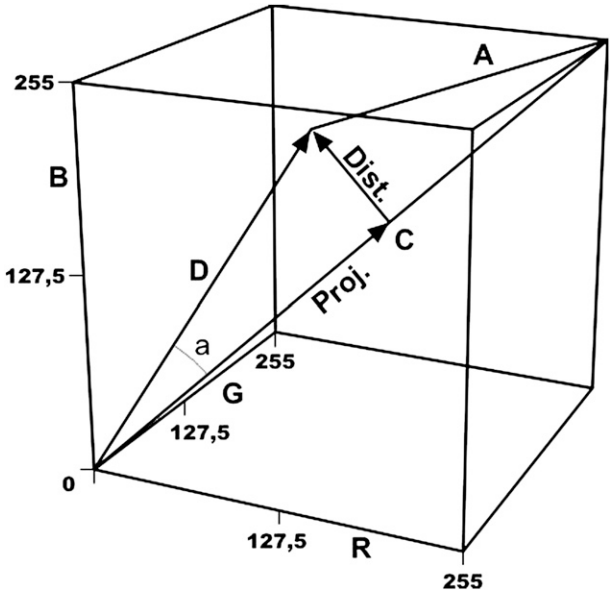

FIG. 3. Typical analysis used on generic pixel determination of EGD on RGB color-space diagonal. Units are in pixel relative intensity.

A discrimination boundary within a 3 standard deviation from the average was established for both patterns to include 95% of pixels in the classification. A summary of boundary values used for sky patterns is illustrated in Table 1. As can be noticed in Table 1, there is an overlapping region of distance limits between 55.30 and 54.01, resulting in error type I (classification of clouds as Rayleigh) or type II (classification of Rayleigh as clouds) according to the pattern analyzed. Type I and II errors are due to misclassification and are caused by the superposition of the end tails of two neighboring Gaussian distributions. A ''real visual boundary'' between cloud and sky patterns in the image is defined here by statistical exclusion. This overlapping region was considered an intermediate pattern, tagged as the ''I'' category classification and colored brown Fig. 6c.

c. *Brightness scale*

Pixel EGPs for cloud or sky patterns were ranked in submultiples of the RGB cube's main diagonal maximum value. These EGP values were divided into six slots, in categories C1–C6 for clouds and R1–R6 for Rayleigh scattering. Categories were selected based on previous work done by the author (Mantelli Neto 2005). This work conceptually considered blue sky as shades of the Rayleigh-scattering effect (Iqbal 1983; Lenoble 1993); clouds as nonselective scattering of solar light in white (Lillesand and Kiefer 1994); and the total simply by sky. EGP values increase from ''darker'' to ''brighter'' according to their brightness value, as shown in Table 2. Shades of green were employed for the sky and shades of red for clouds to avoid color confusion with the original patterns for the results presented in Fig. 6. The discrimination between clouds and Rayleigh scattering was done

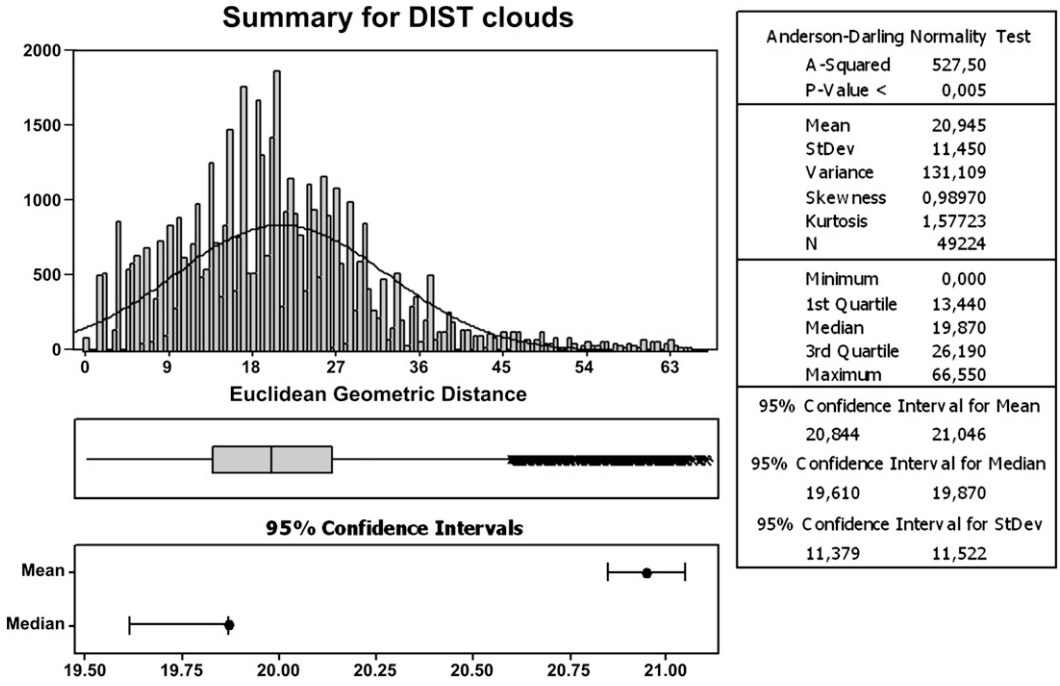

FIG. 4. EDA used for the characterization of the cloud pattern. It includes EGD statistical summary, histogram, box plot, and confidence interval. Horizontal-axis units are in pixel intensity normal to RGB cube's main diagonal.

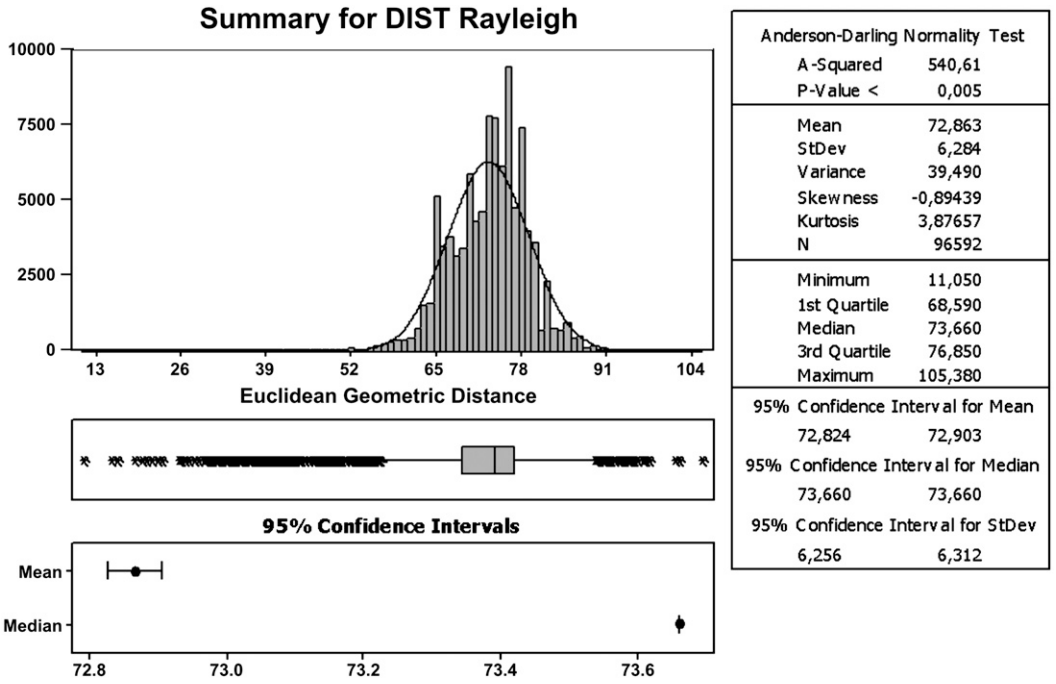

FIG. 5. EDA used for the characterization of the Rayleigh pattern. It includes EGD statistical summary, histogram, box plot, and confidence interval. Horizontal-axis units are in pixel intensity normal to RGB cube's main diagonal.

by statistical methods based on the Gaussian distribution. All the assumptions made on statistical hypothesis tests for additional pixel values were used in the discrimination analysis and have the same meaning. After the characterization of cloud and sky patterns, their typical geometric information was implemented as confidence intervals and loaded into the GUI for pixel classification of other images. Some results of image classification are presented in Fig. 6 and analyzed in the next section.

3. Results

After supervised learning, a set of 49 images relative to one day of observations made on 11 March 2002 every 15 min were analyzed using EGD (the 1900 LT image was missing). Results are presented as the percentage of coverage for each pattern. Images were obtained from output files generated by the GUI tool. A general summary showing pattern data can be observed in Fig. 7, describing the diurnal variability of sky patterns. Data shows dominant categories of pixels mostly between R3 and R6 for sky and C3 and C6 for clouds for the specific image set observed. An increased amount of ''I'' patterns occur mostly in mixed conditions because a threshold (or frontier) between cloud and sky patterns is not formally defined but determined by statistical parameters.

It is important to keep in mind that clouds are not purely white or gray; otherwise, they would have a distribution of points concentrated around the RGB cube's main diagonal or around the HSL cylinder's main axis. The following could explain why they contain some amount of color. Possible colors could be blue due to Rayleigh scattering or red–orange due to Mie scattering. This fact could also be confirmed with additional verification using the HSL color space, as shown in Fig. 8. Rayleigh and Mie effects are present in different sectors of the HSL color-space angle as can be seen in Fig. 8b. For a clear sky, the blue color is limited to a well-defined sector as can be seen in the figure's first row. For a cloudy sky pixel values span one sector with a component of blue and in another sector with red–orange (circled in dashed on second row). For the TSI used in the experiment, there

TABLE 1. Summary of cloud and sky population patterns limits. Units are in pixel relative intensity.

Pattern	Mean	Std dev	N	Q1	Q3	Distance lower limit $\mu - 3\sigma$	Distance upper limit $\mu + 3\sigma$
Cloud	20.945	11.45	49 224	13.44	26.19	0	55.30
Sky	72.863	6.284	96 592	68.59	76.85	54.01	—

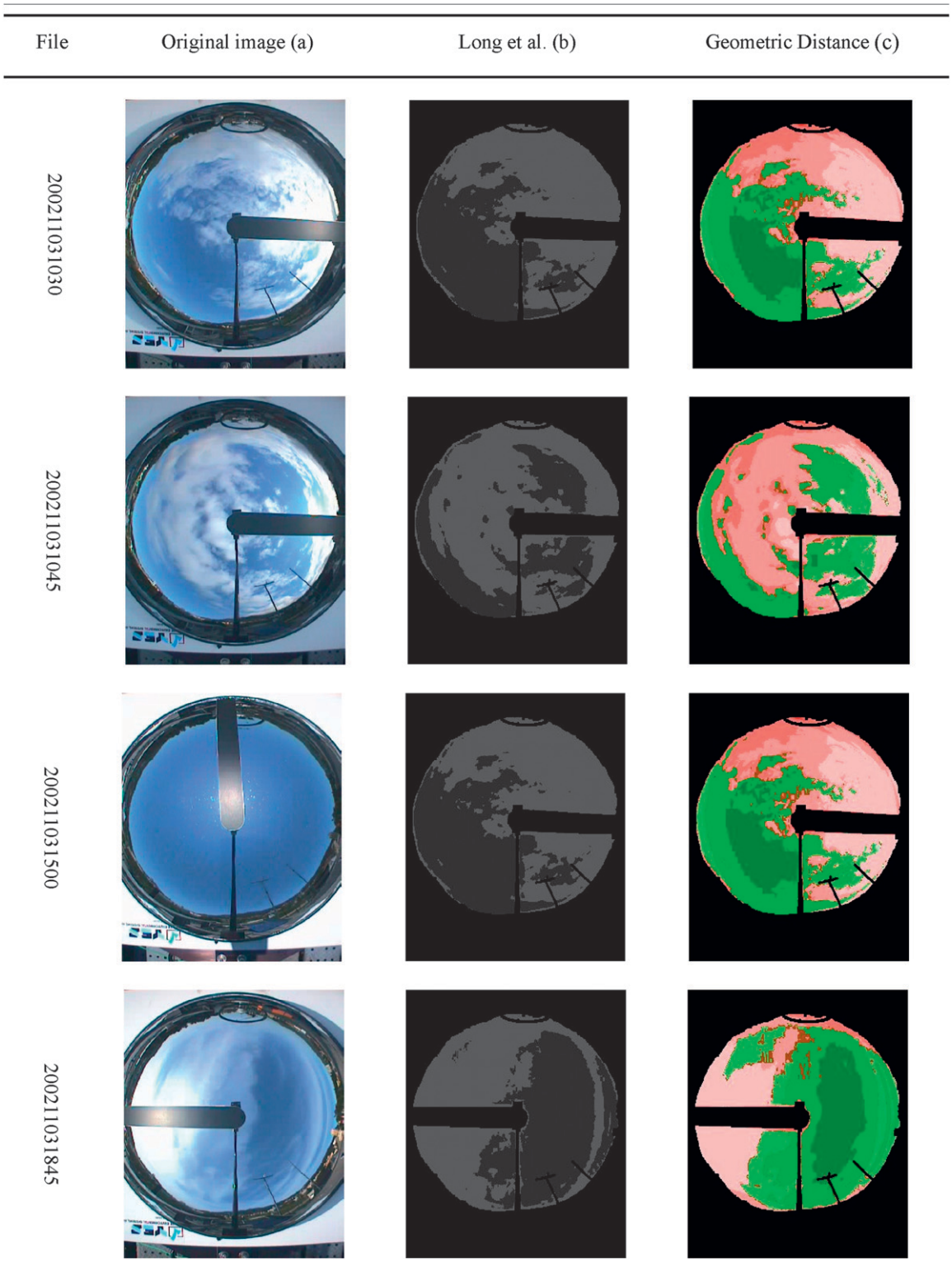

FIG. 6. (a) Original images were compared to (b) images analyzed using Long's method and (c) the geometric distance method. In column (b), clouds were marked in clear gray, and skies were marked in dark gray. In column (c), clouds were colored in shades of red and skies in shades of green. Intermediate patterns were colored in brown in (c).

TABLE 2. Assignments for the pixels' projection values into category slots.

Assigned category	R1 or C1	R2 or C2	R3 or C3	R4 or C4	R5 or C5	R6 or C6
Projection value	0–75	75–150	150–225	225–300	300–375	375–442

are some limitations on Mie-scattering observations because the software takes images only above 5° of solar elevation, thus avoiding the major occurrences of Mie scattering.

Noteworthy here is the robustness of the characterization method when eliminating outliers. Although the elimination of pixels in the middle of the cloud pattern (the second row of Fig. 1) might seem strange, it has an explanation. By using Color Inspector 3D, it was observed that Rayleigh and Mie scattering were intermixed in the original images of clouds, as explained in the previous paragraph. That feature helps to identify and eliminate these patterns from cloud images.

The Rayleigh scattering observed in Fig. 7 shows a diurnal variation with a clearly noticeable gradient in the brightness. It is speculated that this gradient could be associated to other atmospheric contents (aerosols or water vapor), causing diffusion of sun light. For an observer on the surface, Rayleigh scattering is brighter at higher zenith angles than lower ones, indicating that brightness is more intense near the surface (R5) than the lower zenith angles (R3). This is probably due to the diffusion of sun light due to a higher concentration of atmospheric constituents near the surface.

EGD is a distance whose values span from overcast (near the main diagonal) to clear sky, defined here as Rayleigh scattering independent of sun-light variation. This indicates that EGD could also be used to provide information on sky clearness or a similar index. EGD and pixel EGP on the main diagonal (brightness) could be used to support the spatial evaluation of radiation processes from the atmosphere toward the surface. The advantage of the EGD-derived values employed in this analysis is the instantaneous temporal resolution in contrast to indexes obtained from a long temporal series of data. However, further evaluation and careful investigation should be performed to validate and convert pixel values to physical units.

4. Comparisons to related work

There is a limitation in comparing the current methodology with the results of other related works because

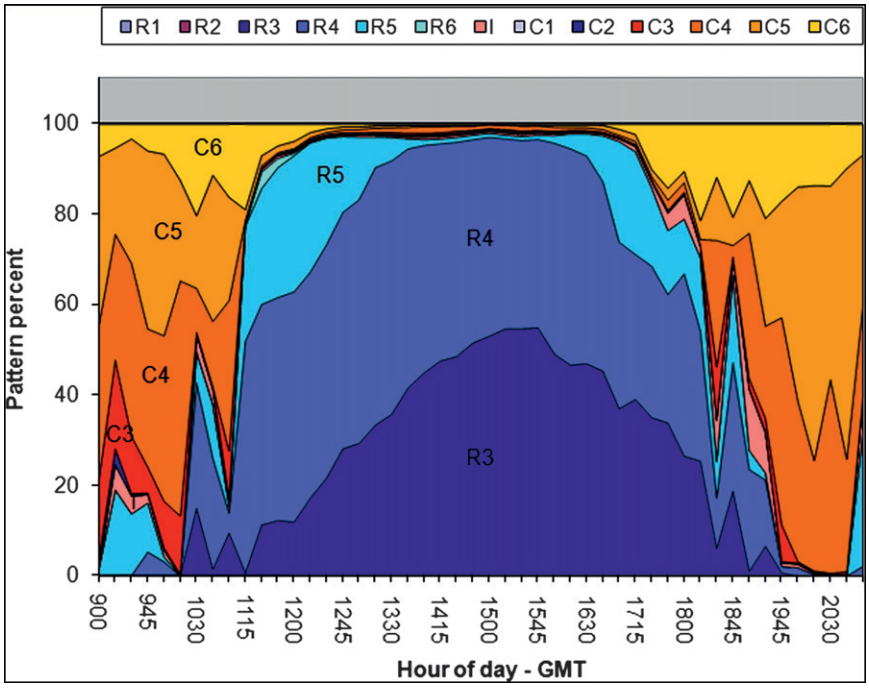

FIG. 7. Sky pattern graph showing the proportion or fraction of sky patterns obtained by the geometric distance method vs time of day. Images were taken from 0900 to 2100 LT every 15 min. The legend categories are in Table 2.

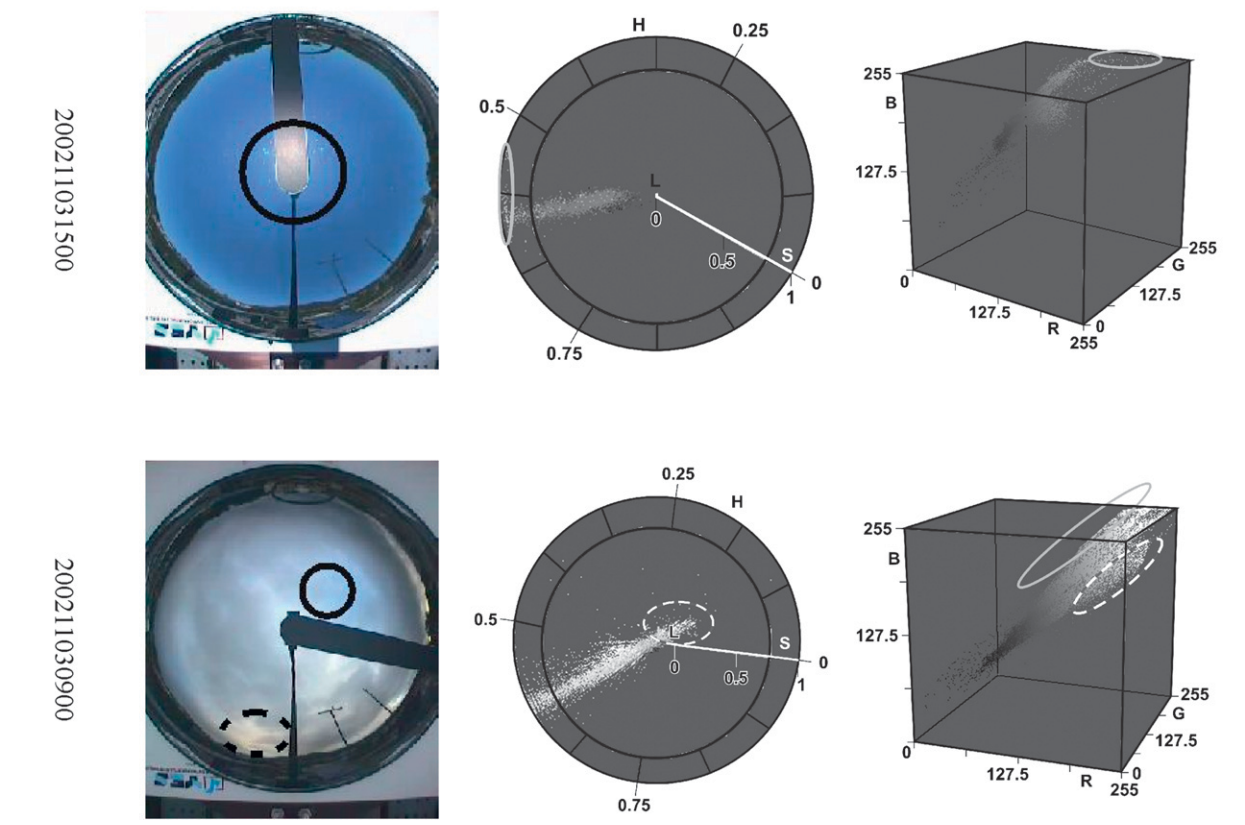

FIG. 8. The first line is used as a comparison showing Rayleigh scattering and the effect of the reflector's surface dust on color space and image, outlined in solid and dashed lines. The second line shows (a) obstruction-masked original images with identification of Mie scattering in dashed lines and Rayleigh scattering in solid lines intermixed in the cloud pattern. (b) The respective HSL color space reduced to *H*, *S* dimensions only showing the presence of Mie scattering in dashed lines is illustrated. Note that the Rayleigh scattering was superimposed by cloud pixels in that perspective view and was not outlined. (c) The respective RGB color space showing typical locus of Rayleigh scattering in solid lines and Mie scattering in dashed lines is illustrated.

of distinct experimental setups and domains used for the desired outcomes.

In Souza-Echer et al. (2006), images were taken from a common camera with observations close to small zenith angles and the sun was always kept out of the FOV. In that condition, pixels representing blue and cloudy skies with high bright values were not analyzed. In high-brightness conditions the saturation values for cloud/sky patterns are mixed with each other, making the discrimination based only on saturation values incomplete. In those conditions, the discriminating function adopted by Souza-Echer et al. (2006) works only on low-brightness pixel values. This is illustrated in Figs. 9a and 9b, where 9a represents an overcast sky and 9b a clear sky. Only the hue and saturation dimensions of the HSL color space are shown, and saturation discriminating values as proposed by Souza-Echer et al. (2006) could be seen as concentric circles. In Fig. 9, the angle *H* represents different colors on the image domain, and the radius is its saturation *S*. High-brightness pixels for sky and clouds are mixed with each other in the same saturation radius, leading to misclassification errors. Therefore, because of the limitation of Souza-Echer et al. (2006) with high-brightness pixels, the classification was not implemented in the GUI tool to be compared with the proposed EGD method.

Long et al.'s (2006) methodology used the same experimental setup as the one presented in this paper but considered only two of the three dimensions available for data analysis. The domain used by Long et al. (2006) is the two-dimensional (2D) coordinates using the red and blue channels and not considering the green channel features. This approach is distinct from the approach used for the

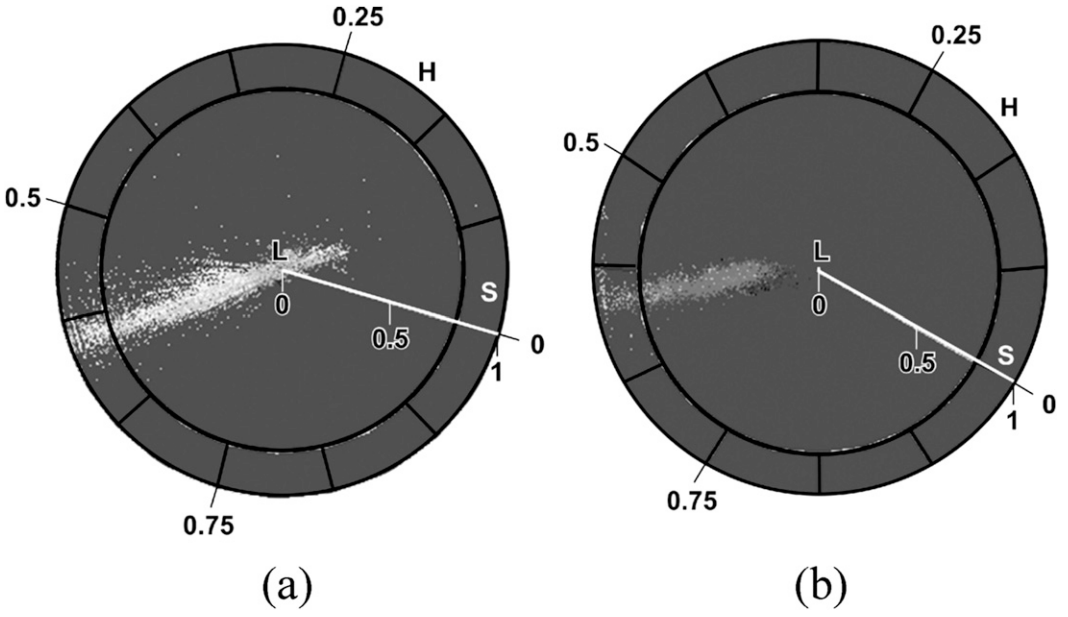

FIG. 9. Differences in the pixel distribution for a cloudy sky from Fig. 1a and a clear sky from Fig. 2b reduced to *S* and *H* dimensions of HSL color space. Circles are placed by the visualization tool and are an example illustrating the type of discrimination function used by Souza-Echer et al. (2006) for each case. Units are in pixel-normalized relative intensity.

present paper, which uses a three-dimensional (3D) domain. We applied the same methodology as Long et al. (2006) to try and reproduce the same technique for the detection of clouds and clear sky. The histograms illustrated in Fig. 10 with the distribution of values are related to the same cloud and sky patterns images of Figs. 1, 2, 4, and 5. As in Long et al.'s (2006) original paper, the same threshold of 0.6 was used for the separation of the cloud and clear-sky patterns.

From that implementation, a selected group of images was analyzed and the comparative results are shown in Table 2. Only a few cases were shown because of space constraints. A complete set of data, including all results and image analyses, can be seen on the Image Processing and Graphics Computing Lab (LAPIX) home page (available online at http://www.lapix.ufsc.br/Clouds/CloudsGeometricDistance/CloudsGeometricDistance.html).

The numeric data obtained from the EGD analysis of 49 images were grouped and reduced to be compared with the Long et al. (2006) method. A summarized proportion or fraction of the sky and cloud pattern along the day comparing Long et al. (2006) and EGD method are shown in Fig. 11.

Figure 12 illustrates a comparative correlation between the two methods for sky and cloud proportions along the day. Pearson's correlation coefficient showed a 0.979 (r^2 = 0.958 441) correlation for cloud proportion and a 0.984 (r^2 = 0.968 256) correlation for sky proportion.

Small differences between the two methods can be explained. The main difference is that Long et al. (2006) established a classification method based on a "reduced dimension empirical discrimination proportion value." The

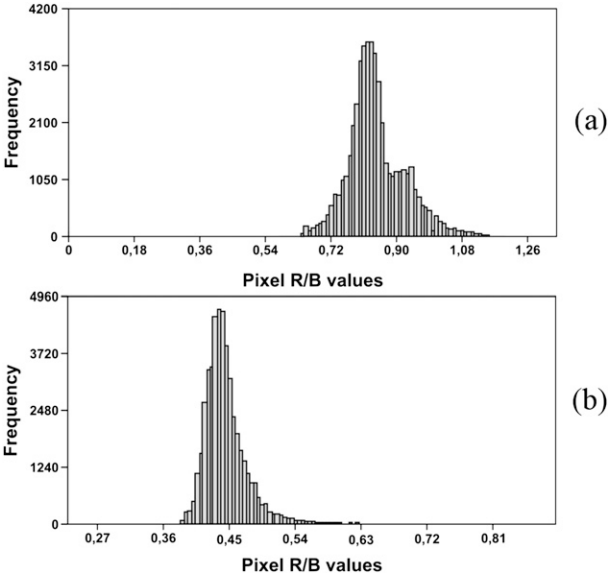

FIG. 10. Histograms of patterns according to the Long et al. (2006) method implemented on the GUI for (a) clouds and (b) clear skies. Horizontal scale is nondimensional.

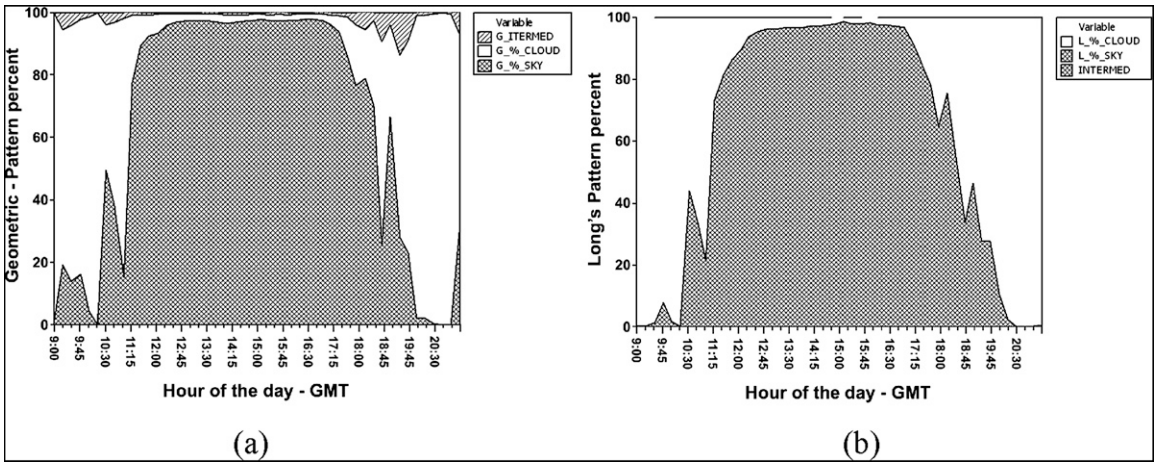

FIG. 11. Comparison of proportions or fraction of sky pattern along the day obtained from (a) geometric distance and (b) Long et al. (2006). Images were taken from 0900 to 2100 at 15-min intervals.

present method is based on "statistics-based confidence intervals." The geometric distance locus considers the tail superposition of statistical characterization due to the already known type I and II errors, not mentioned or considered in the Long et al. (2006) paper. Long et al. (2006) used a hard discrimination criterion, making use of intermediate values classified as either clouds or sky. But for clear-sky and cloud conditions both methods agree. The squared correlation coefficient indicates that the 95.84% variability in clouds and the 96.83% in sky detection in the Long et al. (2006) method are associated with cloud variability in the EGD method. Differences between the two methods are greater, especially when transitory cases are present, because of the occurrence of mixed blue-sky and cloudy conditions in the image. Figure 12 also shows that differences between the two methods tend to increase as the cloud proportion increases.

Looking at the graphics of Fig. 12, a small offset can be observed. It can be inferred that the Long et al. (2006) method assigns more pixels to the cloud proportion than the EGD method. This is confirmed by inspecting the accumulated analysis of both methods throughout the day. Long et al.'s (2006) method indicates a higher amount of cloud proportion and a smaller sky proportion than the EGD method. This explains the observed bias between the two methods. In fact, the determination of

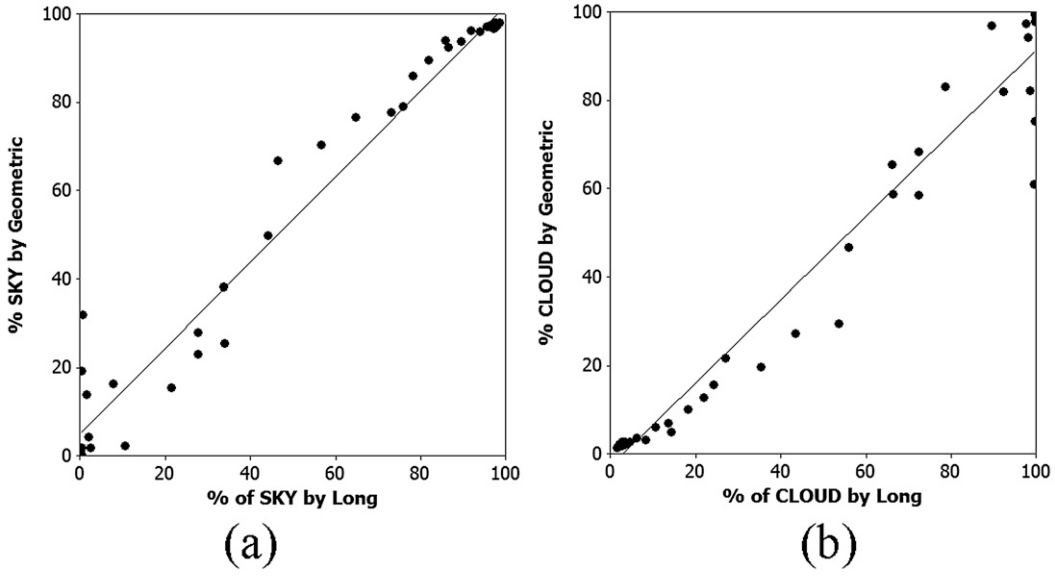

FIG. 12. The comparative correlations between Long et al. (2006) and the EGD methods for (a) sky pattern on and (b) clouds pattern on.

cloud proportions is sensitive to the established discrimination values; as a result, distinct criteria can lead to different results. The differences could be reduced if a characteristic pixel value could be established as the boundary in transition between clouds to sky. Image-preprocessing filtering on smooth transitions like the ones found in the current domain could help to determine that boundary. This is a subject that is recommended to be investigated in future.

5. Future work

Future work based on the Bayesian methodology could explore the use of geometric loci to model other features in the images, especially the ones identifiable by observers but not by automatic systems, which would allow the improvement of qualitative analyses performed by automatic systems. Because of the Rayleigh gradient observed in the sky and its quantification with the brightness EGP scale, we intend to investigate a correlation between the pattern variations to the sun according to photometer measurements. EGD- and EGP-derived values could be correlated to surface radiometers to support the evaluation of a radiative surface flux. In that case, a comparison between EGD's clearness, cloud forcing, and other established time series–based indexes existent in the literature should be investigated. The current methodology is being tested long term at the BSRN São Martinho da Serra station in Southern Brazil.

6. Conclusions

The purpose of the work described in this paper was to develop a methodology to improve the automatic classification of clouds and sky patterns from surface images. EGD allowed not only a classification of those patterns comparable to existing methods but also provided new features in the images based on their color attributes, like Mie scattering and dust removal from the reflector. Those new features could be observed with the 2D (used by Long et al. 2006) and the 1D (used by Souza-Echer et al. 2006) color space–based approaches, because they were masked out by dimensional reduction and mathematical simplifications. The 3D approaches expand the domain analysis of color space to its limit, allowing new potential features to be investigated.

The use of EGD allowed for the classification of clouds and Rayleigh-scattering patterns invariant to their brightness, reducing problems due to solar disk presence, solar variations of the zenith angle, and in the amount of images necessary to model typical occurrences of patterns. Statistical methodologies applied to the image analysis supported those assumptions by means of the Gaussian distribution of patterns and the central limit theorem.

The Bayesian model using supervised learning and analysis applied to EGD patterns proved to be highly correlated to the Long et al. (2006) approach, even when simplified by dimension reduction. This high correlation proved that the Bayesian model used in the present research is a useful tool and could be employed to identify other patterns based on color attributes in future research works. These patterns representing physical phenomena in different color attributes are the same ones perceived by human vision. The mathematical modeling used in the present work matches the theory of the probabilistic model of cognition, as described by Tenenbaum et al. (2006) and Chater et al. (2006), and uses statistical learning and statistical inferences to classify patterns.

Acknowledgments. This work was made possible thanks to the SONDA project sponsored by FINEP (22.01.0569.00) and by PETROBRAS (0050.0029348.07.4). We acknowledge graduate students Leandro Coser, Antonio Carlos Sobieransky, and Adiel Mittmann from LAPIX-UFSC for their support and fruitful discussions on the current subject, and also Professor Sergio Colle from LABSOLAR-EMC-UFSC for providing facilities for the experimental setup.

REFERENCES

Chater, N., J. B. Tenenbaum, and A. Yuille, 2006: Probabilistic models of cognition: Conceptual foundations. *Trends Cognit. Sci.,* **10,** 287–291.

Fernandez-Garcıa, N. L., A. Carmona-Poyato, R. Medina-Carnicer, and F. J. Madrid-Cuevas, 2008: Automatic generation of consensus ground truth for the comparison of edge detection techniques. *Image Vis. Comput.,* **26,** 496–511.

Gonzales, R. C., and R. E. Woods, 2002: *Digital Image Processing*. Prentice Hall, 793 pp.

Holle, L. R., and S. A. Mackay, 1975: Tropical cloudiness from all-sky cameras on Barbados and adjacent Atlantic Ocean area. *J. Appl. Meteor.,* **14,** 1437–1450.

Hoyt, D. V., 1978: Interannual cloud-cover variations in the contiguous United States. *J. Appl. Meteor.,* **17,** 354–357.

Iqbal, M., 1983: *Introduction to Solar Radiation*. Academic Press, 390 pp.

Jiang, X., C. Marti, C. Irniger, and H. Bunke, 2006: Distance measures for image segmentation evaluation. *EURASIP J. Appl. Signal Process.,* **2006,** 209.

Johnson, R. A., and D. W. Wichern, 2007: *Applied Multivariate Statistical Analysis*. 4th ed. Prentice Hall, 773 pp.

Lenoble, J., 1993: *Atmospheric Radiative Transfer*. A. Deepak Publishing, 532 pp.

Lillesand, T. M., and R. W. Kiefer, 1994: *Remote Sensing and Image Interpretation*. John Wiley & Sons, 750 pp.

Long, C. N., D. W. Slater, and T. Tooman, 2001: Total sky imager model 880 status and testing results. ARM Rep. DOE/SC-ARM/TR-006, 36 pp. [Available online at http://www.arm.gov/publications/tech_reports/arm-tr-006.pdf.]

——, J. M. Sabburg, J. Calbó, and D. Pagès, 2006: Retrieving cloud characteristics from ground-based daytime color all-sky images. *J. Atmos. Oceanic Technol.,* **23,** 633–652.

Mantelli Neto, S. L., 2001: Development of a methodology for cloud coverage estimation using surface cameras compared to satellite images. M.Sc. dissertation, Computer Science Department, Federal University of Santa Catarina, 125 pp.

——, A. V. Wangenhein, and E. B. Pereira, 2005: Preliminary model for clouds estimation on RGB color space using automatic imager. *Proc. XII Brazilian Remote Sensing Symp.,* Goiania, Brazil, National Institute of Space Research, 4123–4131. [Available online at http://marte.dpi.inpe.br/col/ltid.inpe.br/sbsr/2004/11.16.17.23/doc/4123.pdf.]

Montgomery, D. C., 2005: *Design and Analysis of Experiments*. John Wiley and Sons, 643 pp.

National Institute of Standards, cited 2010: E-handbook of statistical methods. [Available online at http://www.itl.nist.gov/div898/handbook/.]

Rossow, W. B., 1982: Clouds. *Atlas of Satellite Observations Related to Global change,* R. J. Gurney et al., Eds., International Satellite Cloud Climatology Project, 141–162.

Russell, S., and P. Norvig, 2003: *Artificial Intelligence: A Modern Approach*. 2nd ed. Elsevier, 1021 pp.

Souza-Echer, M. P., E. B. Pereira, L. S. Bins, and M. A. R. Andrade, 2006: A simple method for the assessment of the cloud cover state in high-latitude regions by a ground-based digital camera. *J. Atmos. Oceanic Technol.,* **23,** 437–447.

Tenenbaum, J. B., T. L. Griffiths, and C. Kemp, 2006: Theory-based Bayesian models of inductive learning and reasoning. *Trends Cognit. Sci.,* **10,** 309–318.

World Climate Research Program, 2007: Section of the baseline surface radiation network. WCRP Informal Rep. 9, 54 pp.

World Meteorological Organization, 2008: Guide to meteorological instruments and methods of observations. 7th ed. WMO 8 I.15-1–I.15-11, 681 pp.