ANALYSIS OF EXTREME EVENTS USING A DATA MINING APPROACH

Chapter · January 2015

CITATIONS
0

READS
17

4 authors:

Heloisa Musetti Ruivo
National Institute for Space Research, Brazil
14 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Haroldo Fraga de Campos Velho
National Institute for Space Research, Brazil
406 PUBLICATIONS **1,902** CITATIONS
SEE PROFILE

Fernando Manuel Ramos
National Institute for Space Research, Brazil
204 PUBLICATIONS **1,598** CITATIONS
SEE PROFILE

Saulo R. Freitas
USRA/GESTAR - NASA Goddard Space Flight Center
305 PUBLICATIONS **3,670** CITATIONS
SEE PROFILE

ANALYSIS OF EXTREME EVENTS USING A DATA MINING APPROACH

Heloisa M. Ruivo[1], Haroldo F. de Campos Velho[1], Fernando M. Ramos[1], Saulo R. Freitas[2]

[1]Laboratory for Computing and Applied Mathematics, National Institute for Space Research, São José dos Campos, São Paulo, Brazil
[2]Center for Weather Forecasting and Climate Studies, National Institute for Space Research, São José dos Campos, São Paulo, Brazil

ABSTRACT

The increasing volume of data in the environment sciences is a challenge for analysis and interpretation. Among the difficulties generated by this "data deluge" is the development of efficient knowledge discovery strategies. Here, we apply methods from statistics and computational intelligence to analyze large data sets of climate science. These techniques are simple and robust, and generate a mapping becoming easier the interpretation. Our approach comprises two steps for knowledge extraction. The first step applies a statistical method for class comparison. The second step consists of a Decision Tree (DT) classifier, based on learning algorithm. The DT is used as a predictive model to identify the precipitation intensity. The methodology is used to identify and for understanding extreme rainfall events. The method is employed to identify the more significant meteorological variables associated to the event. The technique is applied to two extreme precipitation events occurred in Brazil: Santa Catarina state (2008), and another one in Rio de Janeiro state (2010).

KEYWORDS

Data mining, statistical analysis, t-test, p-value, decision tree, Shannon entropy.

1. INTRODUCTION

Today, there is scientific evidence that for a warmer planet extreme climate event could become more intense and more frequent [IPCC 2007]. This interpretation has gradually emerged, since the first IPCC Assessment report in 1990. Large amount of data, covering the variables from the atmosphere, land, ice, and ocean, simulations and/or observations at different time intervals and spatial resolutions has been feeding database for recent studies and monitoring. These data sets come from instruments in satellites, *in situ* (ground-based) sensor networks, outputs of computer simulations, and reanalyses [Overpeck 2011]. Among the challenges generated by this deluge of data is the development of better technologies to store, distribute, analyze, and visualize their information content [Hey 2010].

Currently, climatologists have at their disposal a well-established range of statistical tools, from simple and easy-to-use methods of analysis that permit optimal description of sophisticated relationships among extremely large number of degrees of freedom using a few modes or patterns [Kim1999]. However, given the complexities of the climate system and societal concerns regarding the impacts from extreme events, which might be or not related to the climate change, there is still a large demand for the development and use of efficient knowledge discovery techniques.

Data mining (DM), a step in the more general process of Knowledge Discovery in Databases (KDD), extracts information and transform it into an understandable structure for further use, in order to facilitate a better interpretation of existing data [Fayyad 1996]. These patterns can be seen as a kind of summary of the input data and may be used in further analysis. In recent years, DM has been used widely in the areas of science and engineering. Witten and Frank (2000) mention countless applications of machine learning like decision involving judgment, screening images, load forecasting, medical diagnosis, marketing and sales, and so on. In

educational research, Baker (2007) cites the use of DM to study the factors leading students to choose action with impact on reduction on their learning. Such study is applied to understand the factors for university student retention [Baker, 2007].

Mining patterns from Earth Science data is a difficult task due to the spatio-temporal nature of the data. A kind of goal is to discover the spatio-temporal relationship among several climatological variables in the Earth. This is critical for understanding how different variables interact with each other. A standard approach for finding such patterns is to compute the pair-wise correlation between time series of different geographical locations and then, finding regions that have high correlations [Vipin 2001].

Here we present an innovative data mining approach to investigate the climatic causes of extreme events such as the Santa Catarina 2008 flood, and the Rio de Janeiro 2010 flood. Our approach comprises two main steps of knowledge extraction, applied successively in order to reduce the complexity of the original dataset, and identify a much smaller subset of climatic variables that may explain the event being studied. In the first step, we follow along the lines of [Ruivo 2014], and apply a class comparison technique to analyze large data sets. This step results in series of p-value spatial fields that identify which climatic variables behave differently across pre-defined classes of precipitation intensity. The second step consists of a decision tree (DT). It is used as a predictive model to map the set of climate variables with most statistically significant, identified in the previous step. There are many classifiers proposed in the literature (K-means, Bayesian inference, neural networks, and so on). However, the DT allows for producing a hierarchical structure of importance among the different attributes: the first leaf in the DT identifies the most relevant attribute. Indeed, the DT can be employed as a tool to map the relevance of the attributes associate to a specific event. In the present context, the final result identifies a small subset of climatological variables (attributes with smallest p-values) that may explain or even forecast the extreme event.

2. METHODOLOGY

Data mining is a recent technology that potentially identifies the most important information in databases. It is a part of a larger process of KDD. Data mining may be considered as advances in statistical analysis and modeling techniques to find useful patterns and relationships. Generally, DM is the process of analyzing data and summarizing it into useful information.

Data mining algorithms and models such as decision trees, associations, clustering, classification, regression, sequential patterns, and time series forecasting have the potential to identify drought and floods patterns and characteristics. Time series data mining applications organize data as a sequence of events, with each event having a class of occurrence. In data analysis applications on a sequence of events, one of the main challenges is finding similar situations.

In this work, we employ the data mining approach that comprises two steps of knowledge extraction: class-comparison, and decision trees. These methods are applied successively to reduce the complexity of the original dataset and identify a much smaller subset of climatic variables that may explain the event being studied.

2.1. Class-comparison

The class-comparison method is used here for comparing two or more pre-defined classes in a time series of climatic grid box values. The objective is to determine which variables in our data set behave differently across pre-defined classes of precipitation. The "no-difference" case corresponds to a null hypothesis. The classes are defined in such a way so as to capture in the correct class the main episodes of extreme precipitation that occurred during the period being evaluated.

There are several methods for checking whether differences in variable values are statistically significant [Simon 2003]. The F-test is a generalization of the well-known t-test, which measures the distance between two samples in units of standard deviation. Large absolute values of the F-statistic suggest that the observed differences among classes are not due to chance, and that the null hypothesis can therefore be rejected.

Supposing there are J1 data points of class 1 and J2 data points of class 2, the t-test score is computed as:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_p^2\left(\frac{1}{J_1} + \frac{1}{J_2}\right)}}, \qquad (1)$$

$$s_p^2 = \frac{(J_1-1)s_1^2+(J_2-1)s_2^2}{J_1+J_2-2}\ , \qquad (2)$$

$$s_i^2 = \frac{1}{J_{i-1}}\sum_{j=1}^{J_i}\left(x_{ij} - \bar{x_i}\right)^2 \quad (i=1, 2) \qquad (3)$$

where $\bar{x_1}$ and $\bar{x_2}$ are means for samples class-1 and class-2, respectively.

A F-statistic shall be computed for more than two classes. In this case, the alternative to the null hypothesis is that at least one of the classes has a distribution that is different from the others. The t-test and F-test computed are then converted into probabilities, known as p-values. A p-value is the probability that one would observe under the null hypothesis a t-statistic (or F-statistic) as large as or larger than the one computed from the data. Both the t-test and F-test assume that the means are normally distributed, which may not hold, particularly when the number of data points is small.
Therefore, a p-value is the probability of observing an F-statistic as large as or larger than the one computed from the data. It is a measure of statistical significance in the sense that one expects to observe, under the null hypothesis, p-values less than 0.01 only 1% of the time. Permutations methods, which do not rely on data normality assumptions, are commonly used for computing p-values [Simon 2003, Hardim 2007]. After calculating t-test scores for each variable, the class labels of the J1 and J2 are randomly permuted, so that a random J2 of the samples are temporarily labeled as class 1, and the remaining J2 samples are labeled as class 2. Using these temporarily labels, a new t-test score is calculated, say t*. The labels are then reshuffle many times again, with a t* being computed at each permutation. The p-value from the permutation t-test is given by:

$$\mathbf{p-value} = \frac{\mathbf{1+\#of\ random\ permutation\ where\ |t^*|\geq|t|}}{\mathbf{1+\#of\ random\ permutation}}\ . \qquad (4)$$

2.2. Decision tree

There are several decision tree (DT) algorithm available. Here we used the J4.8, a Java implementation of the C4.5 algorithm, from the WEKA package [Witten 2000]. The J4.8 belongs to a succession of DT learners developed by Hunt and others in the late 1950s and early 1960s [Hunt 1962]. DTs are tree-like recursive structures made of leafs, labeled with a class value, and test nodes with two or more outcomes, each linked to a sub-tree.

The DT algorithm construction consists of a collection of training cases, each having a tuple of values for a fixed set of attributes (independent variables) and a class attribute (dependent variable). The aim is to generate a map that relates an attribute value to a given class. The classification task is performed following down from the root the path dictated by the successive test nodes, placed along the tree, until a leaf containing the predicted class.

The analyzed problem is successively divided into smaller sub-problems until each subgroup addresses only one class, or until one of the classes shows a clear majority not justifying further divisions. Most algorithms attempt to build the smallest trees without loss of predictive power. To this end, the J4.8 algorithm relies on a partition heuristic that maximizes the "information gain ratio", the amount of information generated by testing a specific attribute. This approach permits to identify the attributes with the greatest discrimination power among classes, and select those that will generate a tree that is both simple and efficient.

The information gain is measured in terms Shannon's entropy reduction. Given a set A with two classes P and N, the information content (in bits) of a message that identifies the class of a case in A is then

$$I(p, n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right), \qquad (5)$$

where p is the total number of objects belonging to class P, and n is total number of the objects into the classes N. If A is partitioned into subsets A1, A2, … , Av by a given test T, the information gained is given by

$$G(A; T) = I(A) - \sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(A_i)\ , \qquad (6)$$

where Ai has p_i objects from the class P, and n_i from the class N. The algorithm chooses the test T that maximizes the information gain ratio G(A;T)/P(A;T), with

$$\mathbf{P(A;T)} = -\sum_{\mathbf{i=1}}^{\mathbf{v}} \frac{\mathbf{p_i+n_i}}{\mathbf{p+n}} \mathbf{log_2}\left(\frac{\mathbf{p_i+n_i}}{\mathbf{p+n}}\right), \qquad (7)$$

being the information gain from the partition itself. The process is repeated recursively to obtain the other nodes, structuring the decision tree with the rest of the subsets [Quinlan 1993].

3. RESULTS

Class-comparison method is applied to determine which climatological variables in the dataset behave differently across pre-defined classes of precipitation intensity. Decision trees are then used with the climatological variables with smaller values with the aim of generate a map that relates an attribute value to a given class. These methods are applied to reduce the complexity of the original dataset and identify a much smaller subset of climatic variables that may explain the event being studied.

In the figures shown for each study, the p-values at a given grid point can be interpreted as the probability that the observed difference between classes for this variable is the product of mere chance. Clearly, coherent patterns of low p-values (darker areas) attract attention. A p-value < 0.01, for example, indicates probability lower than 1% of being a false positive.

The methodology is employed to the study of climatic causes of two extreme precipitation events: the 2008 flood in Santa Catarina, and the 2010 flood occurred in Rio de Janeiro, Brazil.

3.1 Santa Catarina flood - 2008

The entire data set used in the analysis can be freely downloaded from the Web. Surface- and pressure-level atmospheric fields have a spatial resolution of 2.5º x 2.5º and were extracted from NCEP/NCAR Reanalyzes [Kalnay 1996]. Sea Surface Temperatures (SSTs) on a 2º x 2º grid were obtained from the NOAA Optimum Interpolation SST Analysis, version 2 [Reynolds 2002].

The data set used in this study comprises 3,693 time series as Sea Surface Temperature, Sea Leavel Pressure, Air Surface Temperature, Geopotential height, Cloud cover, Specific Humidity at 850 and 1000 hPa, Omega ($\omega = dp/dt$) at several levels, Meridional and Zonal Wind at 200, 500 and 850 hPa. Gridded data cover a region delimited by latitudes 20ºS and 50ºS, and longitudes 30ºW and 60ºW. Since the episode of extreme rainfall in Santa Catarina was an event of short duration, pentad-averaged anomalies were used in the analysis.

The goal is to identify variables that might correlate with observed differences among classes of precipitation in the region of Blumenau (dot in Figure 1), one of the most affect areas by the 2008 disaster. To this end, we analyzed 12 years (January 1999 up to December 2010) of pentad averages. Precipitation data in the region of Blumenau is an average of five measurement stations of Brazilian National Water Agency (Agência Nacional de Águas, ANA) [SNRH 2010]. Anomalies were computed relative to the mean values over the period 1999-2010.

For classification purposes, the pentads of this time series were divided in three classes of precipitation intensity: “strong”, “moderate”, and “light” rainfall. The standard t-test (eq. 1) was applied, as recommended for applications with two classes: “strong” (precipitation greater than 8), and “moderate” (precipitation between 0 and 8). Results for the most significant variables identified by this procedure are presented in Figure 1. These results represent p-value fields, where coherent spatial patterns of low p-values indicate the existence of a possible links between omega and zonal/meridional wind anomalies, at different levels, and the precipitation intensity in the region of Blumenau. The isolines correspond to omega anomalies averaged over the period November 22 up to 26, 2008, the period of most intense precipitation in Blumenau. The wind fields are also anomalies averaged over the same period.

Regions with darker shades indicate the grid parameters with lower p-values. Figure 1 shows a dense dark area of low p-values for omega at different levels, which extends from the South Atlantic Ocean up the coast of Santa Catarina, and includes in its extreme west the area of Blumenau. This precipitation is fed by moisture transported from the ocean to the continent by easterly winds that predominated in the area in late November.

According to [Dias 2008], the location of a blocking anticyclone on the Atlantic Ocean (with winds that rotate in anti-clockwise on the Southern Hemisphere) determined the occurrence of easterly winds on large part of the South Region coast, resulting in a large scale moisture transport from the ocean to the continent, particularly over the Itajaí valley.

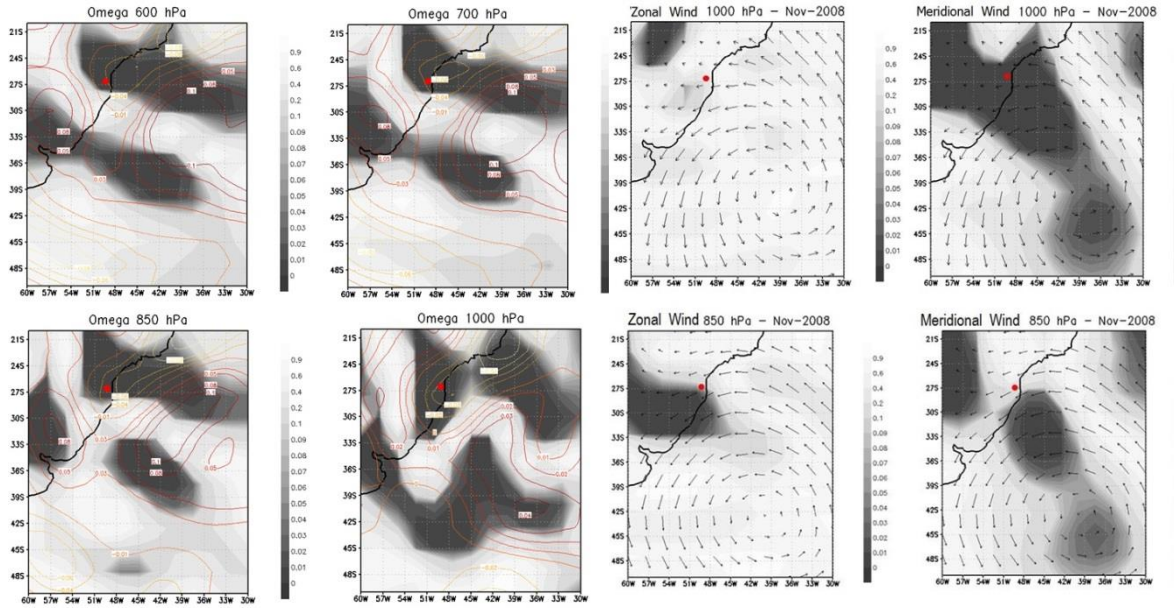

Figure 1. REPRESENTATION in p-values of the climatic variable influence omega (600, 700, 850 and 1000 hPa), and zonal and meridional wind (1000 and 850 hPa) in Santa Catarina flood .

The decision tree with the J4.8 algorithm was created with confidence factor used for pruning (0.25), and number of instances per leaf (8). Several tests were performed: with fixed number of attributes (meteorological variable for different coordinates are considered different attibute) with smallest p-values. The best result was obtained with the 5 different climatological variables, considering 10 different coordinates for each variable, with smallest p-values (total 50 attibutes). To this goal, the precipitation time series were divided over the area of Blumenau in two classes: "light" (values below the median), and "strong" (values above the median), corresponding to episodes of low and high precipitation, respectively. The training set comprised data from 2000 up to 2006. The years of 1999, 2007, 2008, 2009, and 2010 were used to evaluate the tree performance.

The resulting DT, displayed in Figure. 3 right, has 7 leafs (4 "strong" and 3 "light") and 6 decision nodes. The variable with the highest information gain is omega at 500 hPa, and at coordinates 50ºW and 25ºS. As expected, these coordinates are as near to the disaster zone as the limited spatial resolution of the gridded data permits. Note that all but one decision nodes are also associated with omega, at different pressure levels but always in the vicinity of the affected area. These results highlight the key role played in the episode of extreme rainfall in Santa Catarina 2008 by the vertical transport of the moisture, brought from the ocean by sustained easterly winds. As a predictor, the tree was able to forecast 100% of the cases of extreme rainfall during the evaluation years (1999, 2007-2010), including the episode occurred in July 2008.

3.2 Rio de Janeiro flood – 2010

The entire data set used in this study comprises 8,398 time series. Gridded data cover a region delimited by latitudes 21ºS and 24ºS, and longitudes 45ºW and 41ºW. Pentad-averaged anomalies were used in the analysis. Anomalies were computed relative to the mean values over the period 2000-2010 (11 years). Surface- and pressure-level atmospheric fields have a spatial resolution of 0.25º x 0.25º taken to 12 UTC and were extracted from the ECMWF climate reanalysis (www.ecmwf.int/en/forecasts/datasets). ECMWF uses its forecast models

and data assimilation systems to "reanalyze" archived observations, creating global data sets describing the recent history of the atmosphere, land surface, and oceans. The list of the climatic variables is:
- Air temperature at the height level 2 m and pressure levels of 300, 500, 600, 700, 850 925 hPa;
- Geopotential, vertical velocity (omega), specific humidity, zonal and meridional wind components at pressure levels of 300, 500, 600, 700, 850 925 hPa;
- Sea Surface Temperature.

Alerta Rio, a system implemented by the Geotechnical Institute (*Instituto de Geotécnica*) from the Rio de Janeiro city (GeoRio: www.sistema-alerta-rio.com.br), provides rainfall dataset. The precipitation network has 32 sensor stations installed on different places in the city.

For classification purposes, the pentads for this time series were divided in 3 classes of precipitation intensity: "strong", "moderate", and "light" rainfall. The standard t-test (Eq. 1) was applied, as recommended for applications with two classes: "strong" (precipitation greater than 8), and "moderate" (precipitation between 0 and 8). Fields of p-values for eight gridded climatic variables are presented in Figure 2. The isolines correspond to omega anomalies averaged over the pentad April 6 up to 10, 2010, the period of most intense precipitation in Rio de Janeiro. The wind fields are also anomalies averaged over the same period.

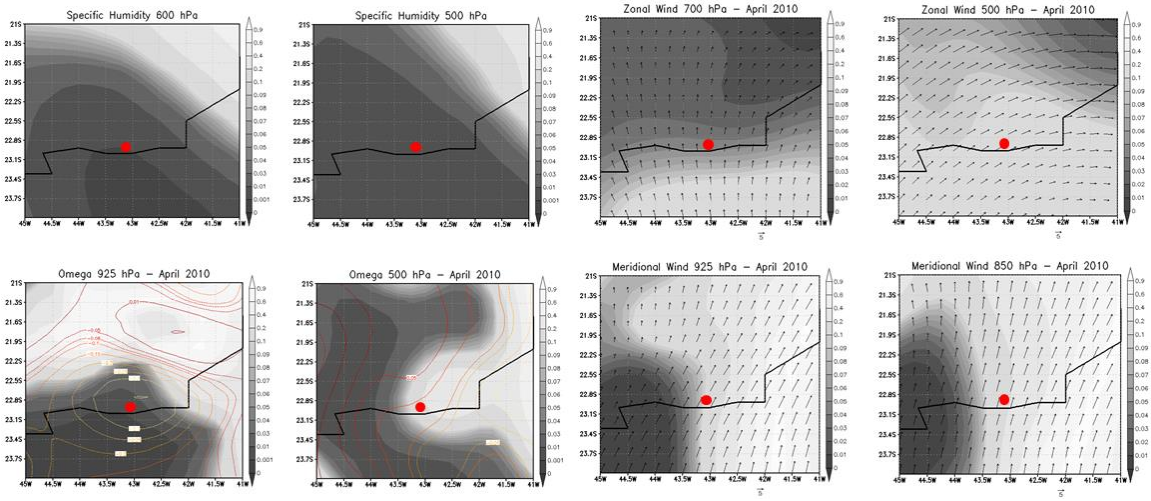

Figure 2. - REPRESENTATION in p-values of the climatic variable influence Specific Humidity (600 and 500 hPa), Omega (925 and 500hPa), zonal wind (700 and 500 hPa), and meridional wind (925 and 850 hPa) in Rio de Janeiro flood .

These results represent p-value fields, where coherent spatial patterns of low p-values indicate the existence of a possible links between specific humidity, omega and zonal/meridional wind anomalies, at different levels, and the precipitation intensity in the region of Rio de Janeiro. Analyzing the Figure 2 there is a dense dark area of low p-values for specific humidity coming from the ocean toward the continent to medium altitudes. It is also observed a dense dark area for omega at 925 hPa on the ocean that spreads to the mainland at 500 hPa. The low p-values in the fields of meridional wind appear in the Southern of Rio de Janeiro state at low altitudes on the other hand, for the zonal wind it is observed low p-values at medium altitudes at the opposite side. The arrows indicate wind transport of moist air from the ocean to the continent.

The decision tree configured by using the J4.8 algorithm was created with confidence factor used for pruning (0.25), with number of instances per leaf equal to 2. Several tests were performed and the best result for the DT was obtained with the 9 different climatological variables, considering 5 different coordinates for each variable, with smallest p-values (total 45 attributes). From the time series for the precipitation anomaly, a DT was configured. The designed DT classifies the Rio de Janeiro precipitation into two classes: "light" (values below 5), and "strong" (values above 5), corresponding to episodes of low and high precipitation, respectively. The training set comprised data annual data from 2000 up to 2006. The years of 2007 to 2010 were used to evaluate the DT performance.

The resulting DT, displayed in Figure 3 left, has 11 leafs (5 “strong” and 6 “light”) and 10 decision nodes. The variable with the highest information gain is omega at 850 hPa (at coordinates 44.5ºW and 23.5ºS). Analyzing only precipitation levels above 5, number of 39 cases was expressed (between 2007 up to 2010), and the DT hits in 13 cases (33.3%). For the considered period (2007 up to 2010), five pentads have rainfall above 5. In these 5 cases, the DT hits the extreme rainfall.

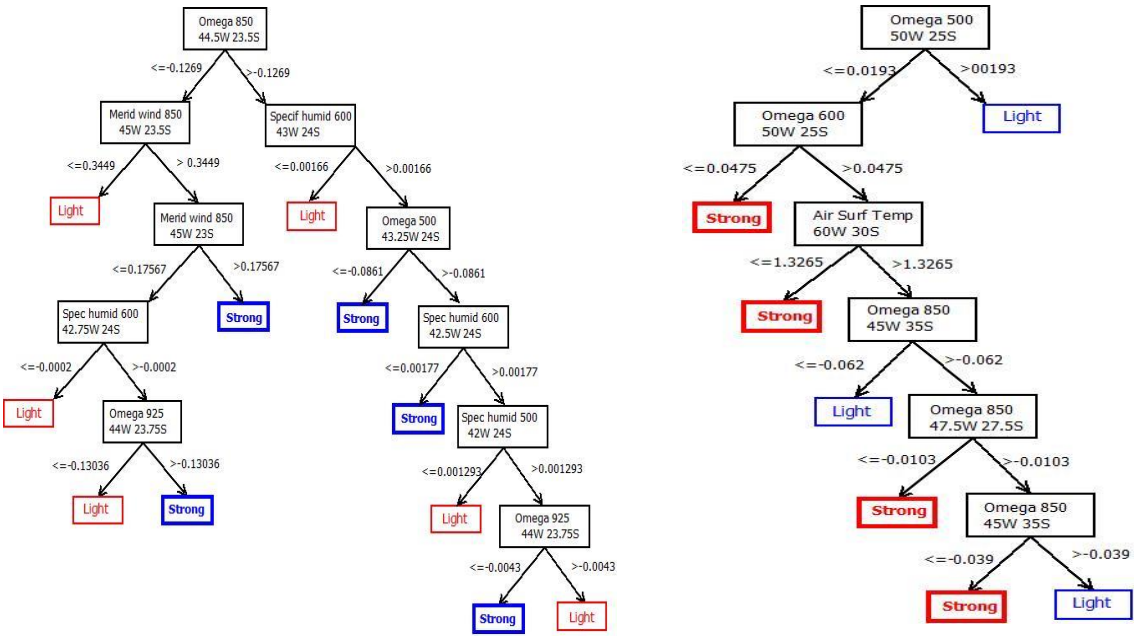

Figure 3: DECISION tree generated: SC flood (right) training set from 2000 up to 2006; test set: 1999, 2007-2010.

RJ flood (left) training set from 2000 up to 2006, and test set: from 2007 up to 2010.

4. CONCLUSION

Currently, given societal concerns regarding the impacts of climate change, there is a growth demand for novel methods, tailored to help with the understanding of climate and its role in the Earth system. In this study, two techniques for data mining were used to investigate the climatic condition behind of two extreme rainfall events occurred in Brazil during the last decade: the Santa Catarina in October 2008, and the Rio de Janeiro, Brazil in April 2010 tragedies.

The class-comparison methodology (p-value evaluation) was able to greatly reduce the size of the original data set, from the order of thousands of variables to a few tenths. The p-values maps becomes easier the interpretation. The p-value analysis also identifies the most relevant climatological variable related to the extreme event. Taking a set with smallest p-values, a decision tree is configured correctly to classify/predict, with high percentage, cases of extreme rainfalls in Santa Catarina state and in Rio de Janeiro city.

The complexity and amount of data today, in particular to the climate science, techniques of data mining and visualization are critical to help us to discover climate patterns, as well as hidden connections among different regions on the Planet. The technique to identify extreme events based on the methodology described and applied in the paper can easily to be implemented in operational centers for weather and climate prediction. The methodology can also be used for analysis of simulated scenarios for climate change studies. The extreme events detection is potentially useful information for the policymakers as well.

ACKNOWLEDGEMENT

Authors thank the Brazilian Agencies for research support (CNPq, FAPESP, CAPES).

REFERENCES

Baker, R. S. J. d., 2007. "Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model". *Workshop on Data Mining for User Modeling 2007*.

Dias M. A. F. S., 2008. *The 2008 November rainfall in Santa Catarina: a case study aimed at improving the monitoring and forecasting of extreme events* [text in Portuhuese]. São José dos Campos: INPE, 2009. 67p

Fayyad U. et al, 1996. *Advances in Knowledge Discovery and Data Mining*. California The MIT Press pp 560.

Foster I., 2006. A two-way street to science's future. *Nature*. Vol 440, pp 419

Hardin J. et al, 2007. A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics*. Vol 8:220. doi={10.1186/1471-2105-8-220}.

Hey T. et al, 2010. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. Available:

http://research.microsoft.com/en-us/collaboration/fourthparadigm/. Accessed 04 Nov 2011.

Hunt, E.B., 1962. *Concept learning:An information processing problem*. New York: Wiley.

IPCC: *Cambio climatico 2007: Informe de sintesis, 2007*. Grupo Intergobernamental de Expertos sobre el Cambio Climáatico [Equipo de redacción principal: Pachauri,R.K. y Reisinger, A. (directores de la publicación), IPCC, inebra, Suiza, 104 págs

Kalnay E. et al, 1996. The NCEP/NCAR 40-Year Reanalyses Project. *Bull Amer Meteor Soc* Vol 77, pages 437-471.

Kim K. Y. and Wu Q., 1999. A comparison study of EOF techniques: Analysis of nonstationary data with periodic statistics, *Journal of Climate*, Vol 21, pp 185-1912, DOI = {http://dx.doi.org/10.1175/1520-0442-12.1.185}}

Overpeck, T.J. et al, 2011. Climate Data Challenges in the 21st Century. *Science* – Vol 331, pp 700-702.

Quinlan J. R., 1993. *C4.5: programs for machine learning.* Morgan Kaufmann Publishers.

Reynolds R.W. et al, 2002. An improved in situ and satellite SST analysis for climate, *Journal Of Climate* Vol 15, pages 1609-1625.

Ruivo, H. M. et al, 2014. Knowledge extraction from large climatological data sets using a genome-wide analysis approach: application to the 2005 and 2010 Amazon droughts. *Climatic Change*; pp 1-15.

Simon R. M. et al., 2003. Design and analysis of DNA microarray investigations. *Springer* Vol 209.

SNIRH -Sistema Nacional de Informações sobre Recursos Hídricos – *Agencia Nacional de Águas (ANA)* - Available in: http://ana.gov.br/portalsnirh/. Accessed March 2010

Vipin K. et al, 2001. Mining Scientific Data: Discovery of Patterns in the Global Climate System. *In Proceedings of the Joint Statistical Meetings*, Athens, GA, Aug. 5–9. American Statistical Association.

Witten I. H. and Frank, E. S., 2000. Data mining: Practical machine learning tools and techniques with java implementation. *Morgan Kaufmann Publishers*.