American Journal of Climate Change, 2018, 7, 420-430
http://www.scirp.org/journal/ajcc
ISSN Online: 2167-9509
ISSN Print: 2167-9495

Data Mining for Flooding Episode in the States of Alagoas and Pernambuco—Brazil

Heloisa Musetti Ruivo[1], Haroldo F. de Campos Velho[1], Fernando M. Ramos[1], Saulo R. Freitas[2,3,4]

[1]Laboratory for Computing and Applied Mathematics, National Institute for Space Research, São José dos Campos, Brazil
[2]Center for Weather Forecasting and Climate Research, National Institute for Space Research, São José dos Campos, Brazil
[3]Goddard Earth Sciences Technology and Research, Universities Space Research Association, Columbia, MD, USA
[4]Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA
Email: helo_mr@hotmail.com

How to cite this paper: Ruivo, H.M., de Campos Velho, H.F., Ramos, F.M. and Freitas, S.R. (2018) Data Mining for Flooding Episode in the States of Alagoas and Pernambuco—Brazil. *American Journal of Climate Change*, **7**, 420-430.
https://doi.org/10.4236/ajcc.2018.73025

Received: December 22, 2017
Accepted: August 20, 2018
Published: August 23, 2018

Abstract

The increasing volume of data in the area of environmental sciences needs analysis and interpretation. Among the challenges generated by this “data deluge”, the development of efficient strategies for the knowledge discovery is an important issue. Here, statistical and tools from computational intelligence are applied to analyze large data sets from meteorology and climate sciences. Our approach allows a geographical mapping of the statistical property to be easily interpreted by meteorologists. Our data analysis comprises two main steps of knowledge extraction, applied successively in order to reduce the complexity from the original data set. The goal is to identify a much smaller subset of climatic variables that might still be able to describe or even predict the probability of occurrence of an extreme event. The first step applies a class comparison technique: *p*-value estimation. The second step consists of a decision tree (DT) configured from the data available and the *p*-value analysis. The DT is used as a predictive model, identifying the most statistically significant climate variables of the precipitation intensity. The methodology is employed to the study the climatic causes of an extreme precipitation events occurred in Alagoas and Pernambuco States (Brazil) at June/2010.

Keywords

Data Mining, Statistical Analysis, T-Test, *p*-Value, Artificial Intelligence, Decision Tree

1. Introduction

The states of Alagoas (AL) and Pernambuco (PE), Brazil, have suffered from 17 to 19 June 2010 a strong flood that caused deaths and several material damages (destruction of roads, bridges, and houses). More than 30 municipalities of the two states declared emergency. From the Damage Assessment Report prepared by the Civil Defence, the tragedy resulted in 24 deaths, 38,030 displaced, 20,962 homeless, and damage and material losses estimated at 971 million dollars. The flood represented a huge loss socio-economic for these poor cities. Figure 1 shows the tragedy location of the tragedy, the destroyed bridge and consequences in buildings located in Palmares (PE) are shown.

Tragedies like these were analysed in previous studies [1] [2] where Data Mining (DM) methodologies were employed. The current paper employs the same methodology used in previous research [1] [2], with focus on data science application in extreme weather events. The previous study used Data Mining (DM) techniques in the analysis of two severe events: deep drought and extreme rainfall.

Drought and heavy rainfall are two examples of "severe weather". However, they are events that differ in time and space scales. Droughts are events characterized for enduring long periods (months or years) and reach a large geographic area. Conversely, heavy rainfall phenomena act over much shorter period, and the breadth of heavy rainfall is much shorter. Thus, the action of the public authority is more critical and necessary in extreme rainfall events, in the sense of mobilizing resources and support actions to the population, requiring almost immediate decision/action (almost "real time") from the decision makers (civil defence).

In the previous paper [3], we have studied the extreme precipitation located in one of the most critical regions in Brazil: Itajai Valley in the State of Santa Catarina. Here, the investigation is performed in the northeast region of the Brazil, where are placed the territories of the states of Pernambuco and Alagoas.

The methodology is implemented for each distinct phenomenon, according to the affected region. In this context, the most relevant climatological variables are identified by DM tools. The variable selection depends on the study area(affected

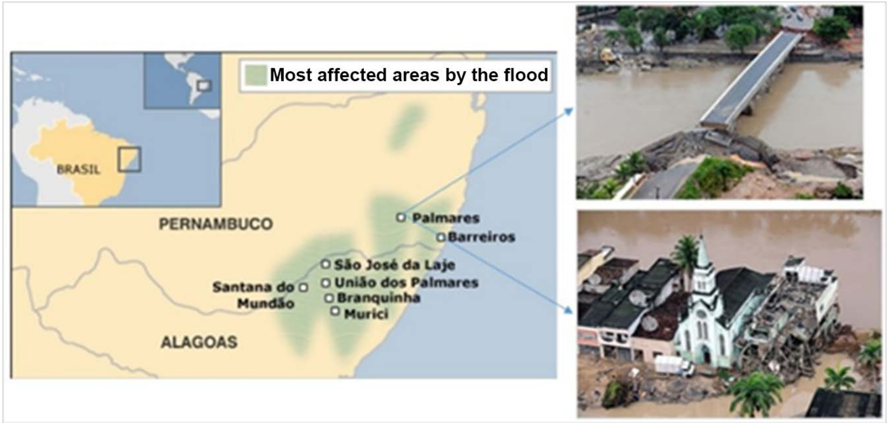


Figure 1. Affected area by the AL-PE flood—adapted from BBC News (Portuguese)—see also accessing the link:
http://www.bbc.com/portuguese/noticias/2010/06/100622chuvasnordesteebc.shtml.

site), type of event (drought or rainfall), and the period or season of the episode [1] [2] [3].

For the analysis of the extreme events above mentioned, DM techniques were used. DM is part of more general process of Knowledge Discovery in Databases (KDD). KDD is the process of analysing data from different perspectives and summarizing it into useful information [4].

Our DM approach comprises two steps of knowledge extraction. The first step uses statistical analysis based on *p*-value computation with two goals: estimate the probability of some attribute linked with the extreme event, and secondly the complexity reduction of the original dataset [1] [2] [3]. The identification of a smaller subset of climatic variables may simplify the understanding of the event under study. The cited statistical evaluation performs a class comparison technique tool to analyse large data set. The *p*-value is computed for different meteorological variable at different location. Therefore, a *p*-value map associated to different variables can show where a certain meteorological value has a stronger link to the event. The second step consists to design of a Decision Tree (DT). The most influential attributes are used as a predictive DT model.

2. Schemes for Data Analysis

Data mining is a recent technology that potentially identifies the most important information in databases. It is apart of a larger process of KDD. Data mining can embraces statistical analysis and modeling techniques to find useful patterns and relationships. Generally, DM is the process of analysing data and summarizing it into useful information. DM algorithms are focused on associations, clustering, classification, regression, sequential patterns, and time series forecasting for many applications. The DM approach comprises the two cited steps of knowledge extraction: class-comparison (*p*-value), and decision tree. 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 and a tool as an event identifier—the Decision Tree.

In these two methods it is necessary to define one bias (threshold) for the calculation of *p*-values, and for the construction of decision trees (more details below). This bias is defined differently in case of rain and drought. Moreover, the parameters used in the heavy rainfall analysis for the region of Santa Catarina in [2] [3] does not apply to the study region (Alagoas, Pernambuco) of Northeast Brazil.

2.1. Statistical Analysis: *p*-Value

The statistical analysis is carried out employing a class-comparison method that compares two or more pre-defined classes of time-series of climatic grid box values. The objective is to determine which variables in the data set behave differently across pre-defined classes of precipitation. There are several methods for checking whether differences in variable values are statistically significant [5].

The F-test is a generalization of the well-known t-test, which measures the distance between two samples in units of standard deviation. T-test can be used to determine if two sets of data are significantly different from each other.

The computed t-test is converted into probabilities, known as *p*-values. The *p*-value is the probability of obtaining a result equal to or "more extreme" than what was actually observed, assuming that the model is true. In this context, the *p*-value is the probability that one would observe under the null hypothesis a t-test as large as or larger than the one computed from the data.

Permutations methods, not making Gaussian assumptions, are commonly used for computing *p*-values [5] [6]. After calculating t-test scores for each variable, the class labels of different classes are randomly permuted. So, considering two classes J_1 and J_2, a random J_2 of the samples are temporarily labelled as class 1, and the remaining J_2 samples are labelled as class 2. Using these temporary 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 computed.

If the *p*-value is smaller or equal than a threshold, then the assumption is acceptable, otherwise, if the *p*-value is greater than the threshold, then assumption can be considered false. The user defines the threshold. Therefore, small *p*-values are linked with larger statistical significance.

2.2. Classifier from the Artificial Intelligence

There are several classifiers based on Artificial Intelligence (AI). The DT classifier is a "divide-to-conquer" approach to the problem of learning from a set of independent instances, leading naturally to a style of data representation [7]. DTs are tree-like recursive structures made of leafs, labelled 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.

Among the free approachable algorithms, here the J4.8 algorithm is applied, which is WEKA's implementation of the decision tree learner [7] [8]. 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 allows attributes identification 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. The quantities of the form $H(S)$ play a central role in information theory as measures of information, choice and uncertainty [7] [8] [9] [10]:

$$H(S) = -\sum\nolimits_{x \in X} p(x)\log\{p(x)\} \tag{1}$$

where S is a given data set, x is the set of classes in S, $p(x)$ is the proportion of the number of elements in class x to the number of elements in S. Information Gain (IG) is the measure of the difference in entropy from before to after the data set S is split on an attribute A [11]:

$$IG(A,S) = H(S) - \sum_{t\in T} p(t)H(t) \tag{2}$$

Being T is the subset created from splitting set S by attribute A such that $S = U_{t\in T^t}$.

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 configured using climatological variables with the smallest p-values. The DT aim is of generate a map relating an attribute value to a given class. Coherent patterns, meaning low p-values (darker areas), indicate high probability for the attributes to be associated with the meteorological event under consideration—in our study: intense rainfall.

The data set used in this study comprises 16,530 time series of surface- and pressure-level atmospheric field with spatial resolution of 0.25˚ × 0.25˚. The dataset were extracted from ECMWF [12] at 12UTC. The climatological variables of surface used in the analysis were: sea surface temperature, and geopotential height. Completing the database, the following variables were used: air temperature at 2 m, and air temperature, specific humidity, omega, meridional and zonal wind at pressure levels: 925 hPa, 850 hPa, 700 hPa, 600 hPa, 500 hPa, 300 hPa. Gridded data cover a region delimited by latitudes 11˚S and 6˚S, and longitudes 33˚ W and 38˚ W. Since the episode of extreme rainfall is an event of duration from one hour up to some days, pentad-averaged anomalies (average on 5 days) were used over the period January 2000 up to December 2010.

Precipitation data is an average of five measurement stations from the National Institute of Meteorology (INMet: Instituto Nacional de Meteorologia—Brasilia (DF), Brazil): Surubim (PE), Arcoverde (PE), Garanhus (PE), Recife Curado (PE); Palmeira dos Indios (AL). The precipitation pentad anomaly series is shown in **Figure 2**. Peaks of precipitation close to 15 mm or greater can be considered as extreme events.

Results with Statistical Analysis: *p*-Value

For classification purposes, the precipitation series is divided into classes. The pentads of the time series were divided in three classes of precipitation intensity: strong, moderate, and light rainfall. The standard t-test was applied, as recommended for applications with two classes: "strong" (precipitation greater than 5), and "moderate" (precipitation between 0 and 5). The red dot points in figures indicate de location of the INMet stations.

Regions with darker shades indicate the grid parameters with lower p-values. **Figure 3** shows a dense dark area of low p-values for air temperature at 2 m and

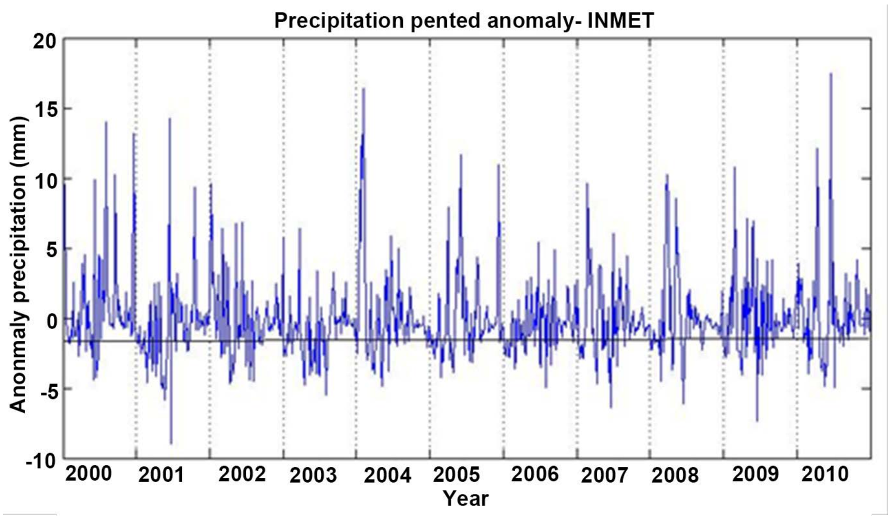


Figure 2. Precipitation pentad anomaly (INMET) at 12UTC—average 5 seasons—2000-2010.

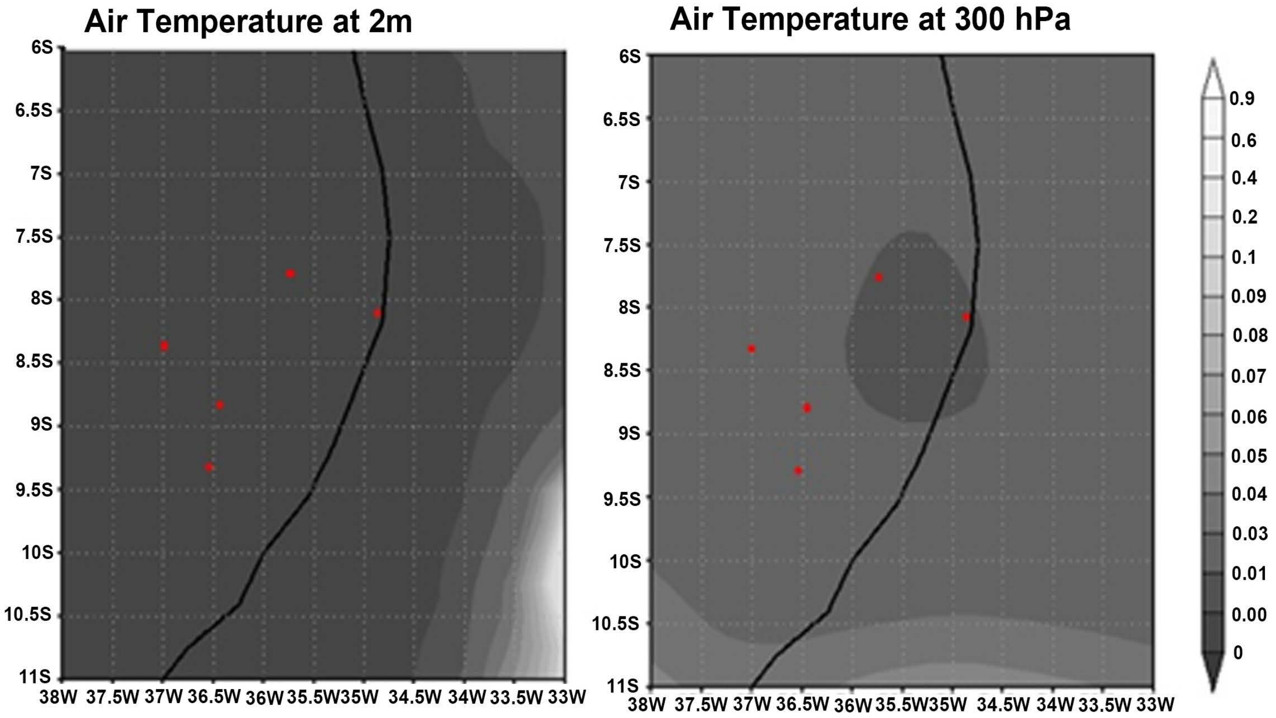
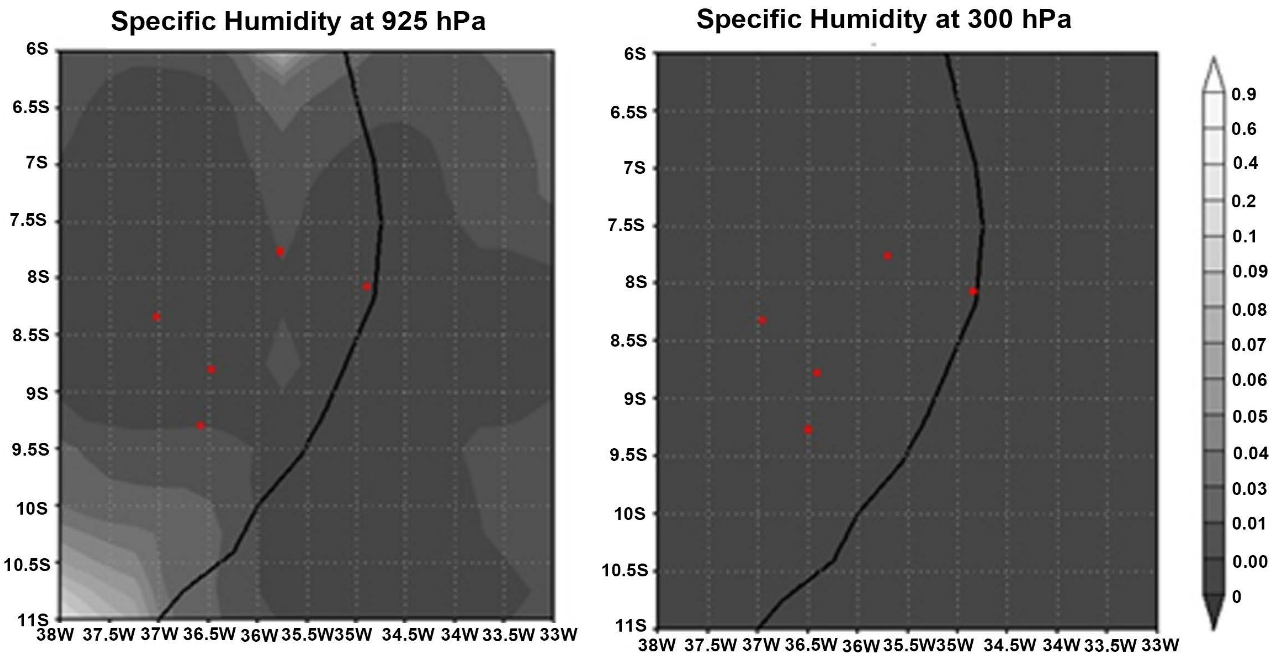


Figure 3. p-values field for air temperature at 2 m and 925 hPa in AL-PE flood.

300 hPa. Low p-value area of Specific Humidity at 925 hPa and 300 hPa is displayed in Figure 4.

Very cold temperature on the top of clouds (Air Temperature at 300 hPa) and high moisture were cited in [13] as relevant factors for precipitation. It was also observed high values of sea surface temperatures in most of the Equatorial Atlantic, particularly near the coast of the Northeast (NE) Brazilian region, contributing with the intensification of the convergence of moisture flux over the coast [13] [14]. This can be seen in **Figure 5** that shows low p-values in sea


Specific Humidity at 925 hPa

Specific Humidity at 300 hPa

Figure 4. *p*-values field for specific humidity 925 and 300 hPain AL-PE flood.

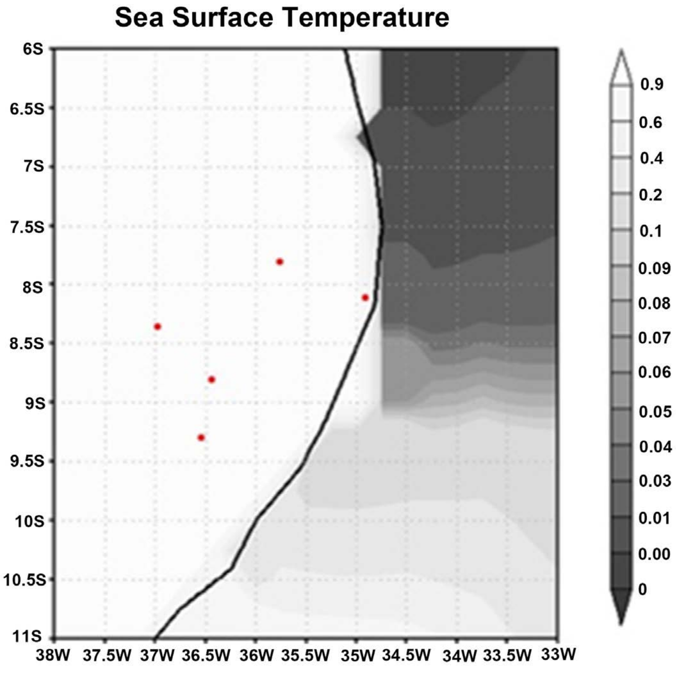

Sea Surface Temperature

Figure 5. *p*-values field for sea surface temperature in AL-PE flood.

surface temperature in the North Atlantic. It can be seen in Figure 6 low *p*-values of omega (upward motion) for both the affected area and the North Atlantic Ocean.

Figure 7 shows a low *p*-value area of zonal and meridional wind on North part of the map. The wind fields correspond to anomalies averaged over the June 15 up to 19, 2010, the period of most intense precipitation.

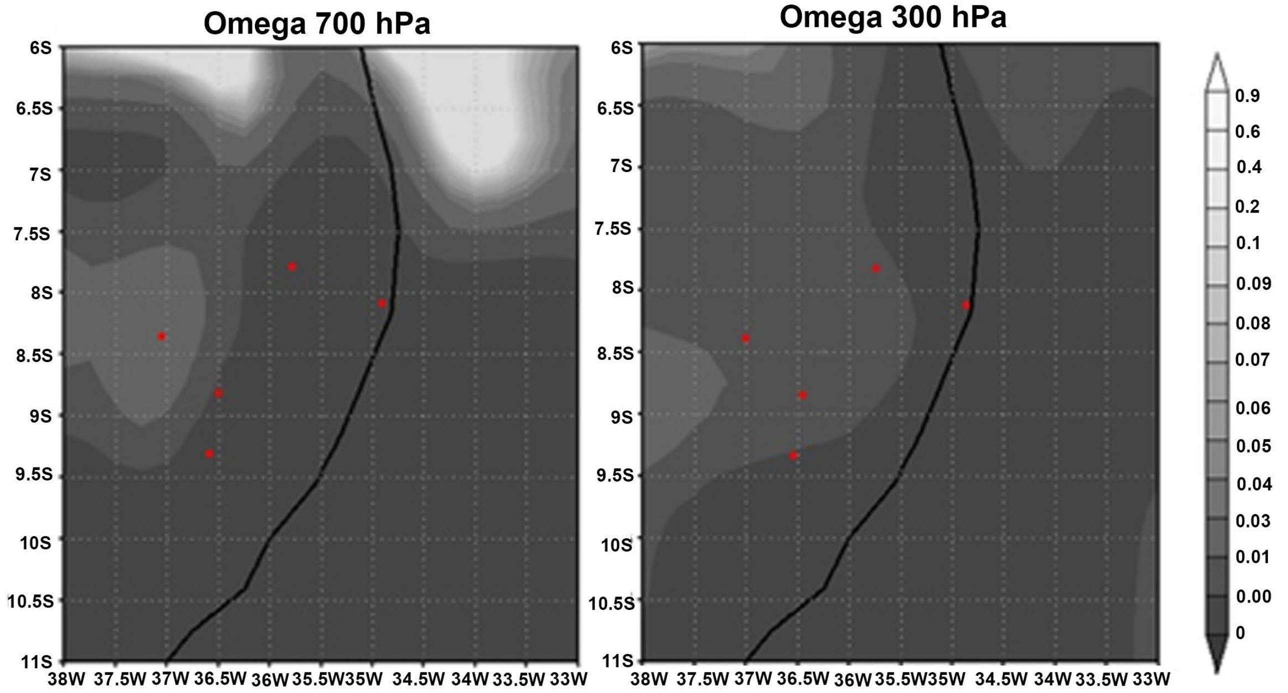


Figure 6. *p*-values field for Omega at 700 and 300 hPain AL-PE flood.

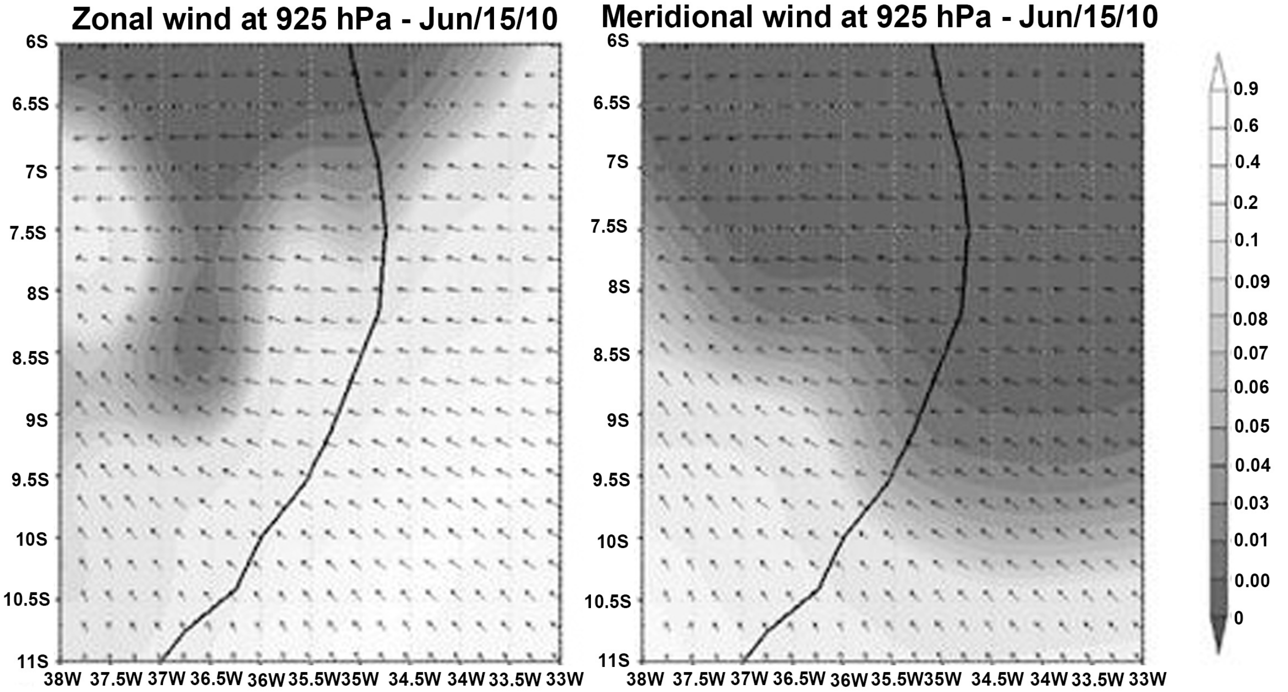

Zonal wind at 925 hPa - Jun/15/10
Meridional wind at 925 hPa - Jun/15/10
6S
6.5S
7S
7.5S
8S
8.5S
9S
9.5S
10S
10.5S
11S
38W 37.5W 37W 36.5W 36W 35.5W 35W 34.5W 34W 33.5W 33W
0.9
0.6
0.4
0.2
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.01
0.00
0

Figure 7. *p*-values field for zonal and meridional winds at 925 hPain AL-PE flood.

The decision tree was designed with the help of J4.8 algorithm. The DT configuration was computed with the 16 different climatological variables, considering 5 different coordinates for each variable, with smallest *p*-values, performing 80 attributes. For the DT classifier, the precipitation time series were divided in two classes: "light" (values below 3), and "strong" (values above 3). The training set comprised data from the year 2000 up to 2006. The years from 2007 up to 2010 were used to evaluate the DT performance. As a predictor, the DT was able to identify the extreme rainfall occurred in 2010 July (Figure 8).

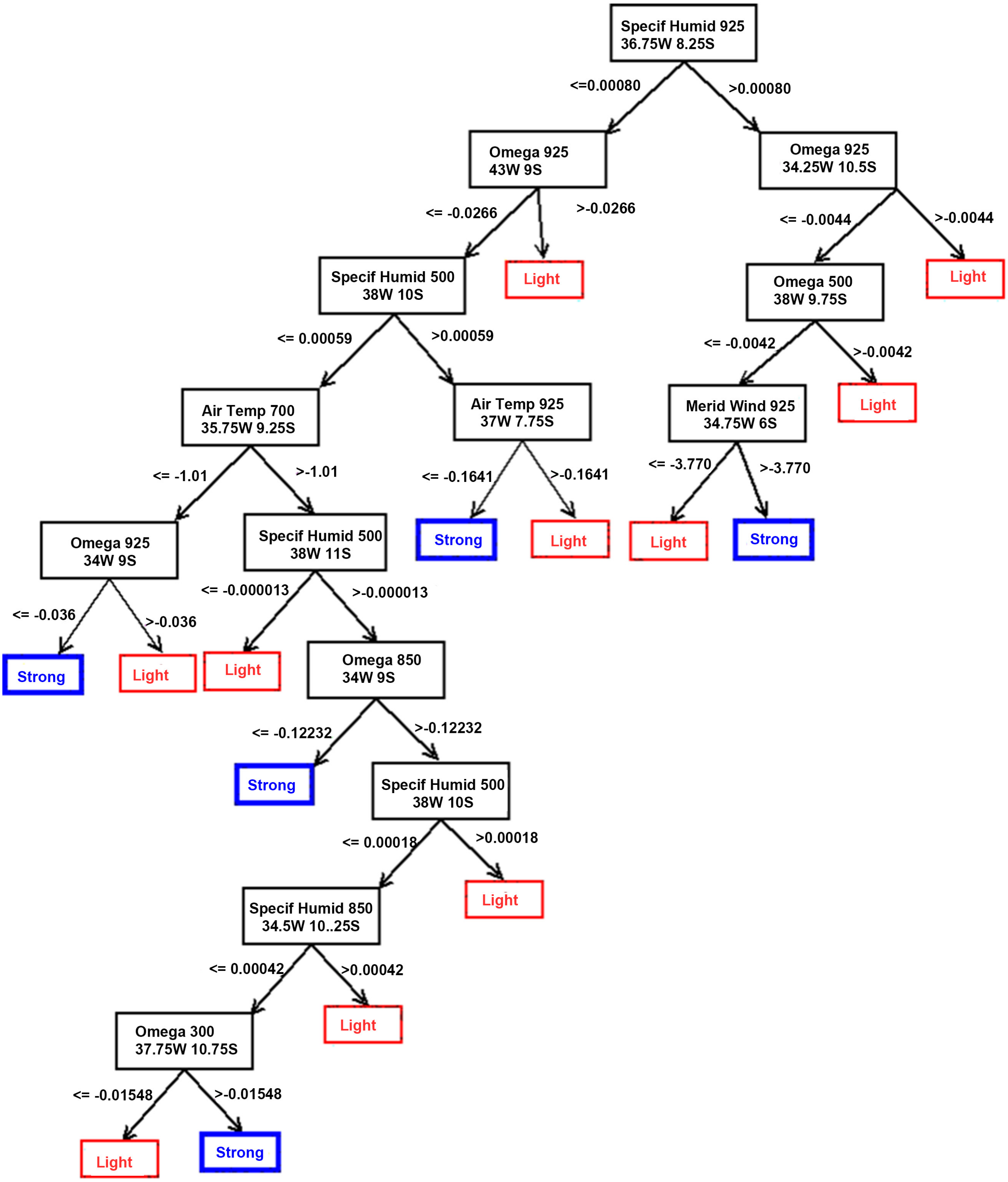


Figure 8. Decision Tree (DT) using training set from 2000 up to 2006, and test set: from 2007 up to 2010.

4. Conclusions

In this study, two techniques for data mining were used to investigate the climatic variables linked to the extreme rainfall events which occurred in Pernambuco and Alagoas States (Brazil) in the year 2010. The class-comparison methodology (*p*-value) was able to identify the most relevant variables. In addition, the scheme was also significant for size reduction of the original data set: from the order of thousands of attributes to a few tenths. A geographical mappinf of *p*-values become the interpretation an easier way for specialists, pointing out climatological variable directly related to the extreme event. The decision trees trained with the reduced dataset computed from the results of the class-comparison previous step were able to correctly classify the cases of extreme rainfall in 2010 for the analysed region in the Brazil.

Overall, the applied data mining procedure has shown to be a promising approach in the investigation of climatic extreme events and the extraction of knowledge from large and complex data sets.

Acknowledgements

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

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ruivo, H.M., Sampaio, G. and Ramos, F.M. (2014) Knowledge Extraction from Large Climatological Data Sets Using a Genome-Wide Analysis Approach: Application to the 2005 and 2010 Amazon Droughts. *Climatic Change*, **124**, 347-361. https://doi.org/10.1007/s10584-014-1066-7

[2] Ruivo, H.M., Campos Velho, H.F., Sampaio, G. and Ramos, F.M. (2015) Analysis of Extreme Precipitation Events Using a Novel Data Mining Approach. *American Journal of Environmental Engineering*, **5**, 96-105.

[3] Ruivo, H.M., Campos Velho, H.F., Ramos, F.M. and Sampio, G. (2013) P-Value and Decision Tree for Analysis of Extreme Rainfall. *Ciência e Natura*, **1**, 231-234. https://doi.org/10.5902/2179460X11604

[4] Fayyad, U., Piatesky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996) Advances in Knowledge Discovery and Data Mining. The MIT Press, Cambridge.

[5] Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W. and Zhao, Y. (2003) Design and Analysis of DNA Microarray Investigations. Series: Statistics for Biology and Health, Vol. 209, Springer, Berlin.

[6] Hardin, J., Mitani, A., Hicks, L. and VanKoten, B. (2007) A Robust Measure of Correlation between Two Genes on a Microarray. *BMC Bioinformatics*, **8**, 220. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-220

[7] Witten, I.H. and Frank, E.S. (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation. 2nd Edition, Morgan Kaufmann Pub-

lishers, Burlington.

[8] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Burlington.

[9] Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 623-656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

[10] Shannon, C.E. and Weaver, W. (1949) The Mathematical Theory of Communication. University of Illinois Press, Champaign.

[11] Mitchell, T.M. (1997) Machine Learning. The Mc-Graw-Hill Companies, New York.

[12] Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., *et al.* (2011) The ERA-Interim Reanalysis: Configuration and Performance of the Data Assimilation System. *Quarterly Journal of the Royal Meteorological Society*, **137**, 553-597.

[13] Fialho, W.M.B. and Molion, L.C.B (2011) Eventos Extremos: Alagoas Junho de 2010. UFPel, Pelotas.

[14] Climanálise (2010) Boletim de Monitoramento e Análise Climática. CPTEC/INPE, Vol. 25. http://climanalise.cptec.inpe.br/~rclimanl/boletim/pdf/pdf10/jun10.pdf