Fractal-based Analysis to Identify Trend Changes in Multiple Climate Time Series

Santiago Augusto Nunes[1], Luciana A. S. Romani[2], Ana M. H. Avila[3], Caetano Traina Jr[1], Elaine P. M. de Sousa[1], Agma J. M. Traina[1]

[1] ICMC - University of São Paulo at São Carlos - Brazil
bynael@grad.icmc.usp.br, {caetano, parros, agma}@icmc.usp.br
[2] Embrapa Agriculture Informatics at Campinas - Brazil
luciana@cnptia.embrapa.br
[3] Cepagri - State University of Campinas - Brazil
avila@cpa.unicamp.br

Abstract. In the last few decades, huge amounts of climate data have been gathered and stored by several institutions. The analysis of these data has become an important task due to worldwide climate changes and the consequent social and economic effects. In this work, we propose an approach to analyzing multiple climate time series in order to identify intrinsic temporal patterns and trend changes. By dealing with multiple time series as multidimensional data streams and combining fractal-based analysis with clustering, we can integrate different climate variables and discover general behavior changes over time.

Categories and Subject Descriptors: H. Information Systems [**H.2 Database Management**]: H.2.8 Database Applications—*Data Mining*

Keywords: anomalies, data streams, clustering

1. INTRODUCTION

In the last decades, results from Climatology have shown worldwide climate changes, mainly with rising temperatures and changing rainfall distribution. According to the World Meteorological Organization, climate change can be defined as the average description of weather conditions inferred from continuous observations during at least a 30-year period [Zhai et al. 2005]. Moreover, deviations from the average description characterizing natural variability and extreme phenomena must also be considered.

Climate research has been carried out to study the influence of climate conditions on global and regional populations and their daily lives, geographic distribution and economic activities such as agriculture. Recently, this research has been intensified due to the increasing world population [Ayoade 1996] and the huge volumes of data gathered from different sources, such as meteorological sensors, weather satellites and climate models. By analyzing such data, meteorologists aim to understand extreme conditions and climate anomalies.

Results from several scientific analyses show an increase in intensity-duration-frequency of extreme events [Alexander et al. 2006] and a consequent escalation of natural hazards. Intense rainfall in a single day as well as consecutive days of precipitation may cause floods and serious problems for both urban and rural areas. Therefore, understanding trends of extreme phenomena is crucial to prepare for adverse situations, i.e., to create conditions to mitigate some of the problems and to make strategic decisions in a feasible time.

Climate-related observations from ground-based meteorological stations, remote sensors, weather radars and other sensors have continually generated a huge volume of data. Furthermore, data from climate models have enlarged climate archives in the magnitude of terabytes per simulation of climate change scenarios. Thus, the analysis of these data has become increasingly challenging for researchers

from several scientific areas. Although well-known statistical methods such as principal component analysis have been widely applied, the complexity and the volume of available data, as well as the need of precise and quick answers, have challenged the scientists. Thus, specialists are motivated to try new methods to retrieve relevant information and discover interesting patterns from climate datasets, also considering correlation identification and integrated analysis of multiple, long time series. In this context, the fractal theory appears as a feasible approach to support two relevant tasks:

(1) efficient analysis of multiple climate time series to find patterns and trend changes;
(2) identification of extreme climate events that indicate regional climate changes.

This article presents a process for climate time series analysis based on concepts from the Fractal Theory. Our approach deals with multiple time series as a multidimensional data stream, such that each time series defines an attribute of the stream. Therefore, it is possible to integrate multiple climate variables in a unified analysis process. In particular, we combine:

(1) fractal data stream monitoring for pattern discovery and trend change detection considering the intrinsic correlation among time series defined by different climate variables;
(2) and clustering to find similar (or distinct) patterns revealed when data are analyzed with different temporal granularities.

Experimental studies carried out on real and estimated climate time series indicate that our approach can be a useful tool to assist specialists in analyzing large amounts of climate data.

The rest of this article is organized as follows. Section 2 presents background concepts of the Fractal Theory and their application on data stream analysis. Section 3 describes our approach to analyze climate time series. Experimental results are discussed in Section 4. Finally, Section 5 presents final remarks and future work.

2. BACKGROUND

A fractal can be defined by the self-similarity property, i.e., an object that presents roughly the same characteristics over a large range of scales [Schroeder 1991]. Accordingly, a real dataset exhibiting fractal behavior is exactly or statistically self-similar, such that parts of any size of the data present the same characteristics of the whole dataset.

From the Fractal Theory, the fractal dimension is particularly useful to data analysis, as it provides an estimate of the intrinsic dimension D of real datasets. The intrinsic dimension gives the dimensionality of the object represented by the data regardless of the dimension E of the space in which it is embedded. In other words, D measures the non-uniformity behavior of real data [Faloutsos and Kamel 1994; Traina et al. 2005]. For instance, a set of points defining a plane embedded in a three-dimensional space ($E = 3$) has two independent attributes and a third one correlated to the others, resulting in $D = 2$.

The fractal dimension of real datasets can be determined by the Correlation Fractal Dimension D_2. An efficient approach to measure the fractal dimension of datasets embedded in E-dimensional spaces is the *Box-Counting* method [Schroeder 1991], which defines D_2 as presented in Equation 1, where r is the side of the cells in a (hyper) cubic grid that divides the address space of the dataset and $C_{r,i}$ is the count of points in the ith cell.

$$D_2 \equiv \frac{\partial log(\sum_i C_{r,i}^2)}{\partial log(r)} \qquad r \in [r_1, r_2] \tag{1}$$

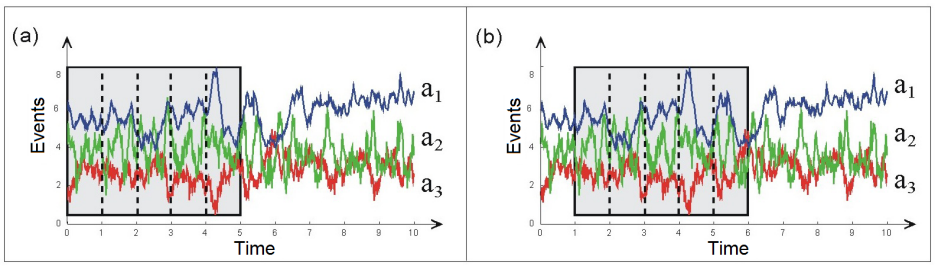

Fig. 1. Sliding window over a three-dimensional *data stream*.

An efficient algorithm (linear cost on the number of elements in the dataset) to compute D_2 was proposed by Traina et al. [2000]. Thus, D_2 can be a useful tool to estimate the intrinsic dimension D of real datasets with feasible computational cost.

Concepts from the Fractal Theory have been applied to several tasks in data mining and data analysis, such as selectivity estimation [Baioco et al. 2007], clustering [Barbará and Chen 2000], time series forecasting [Chakrabarti and Faloutsos 2002], correlation detection [Sousa et al. 2007] and data distribution analysis [Traina et al. 2005].

The information of intrinsic behavior provided by the fractal dimension D_2 can also be applied to detect behavior changes in evolving data streams. Essentially, the idea is to continually measure the fractal dimension of the data stream over time in order to monitor its evolving behavior. Thus, significant variations in successive measures of D_2 can indicate changes in the intrinsic characteristics of the data.

Sousa et al. [2007] proposed a technique to track behavior changes of evolving data streams and the algorithm *SID-meter* to continually measure D_2 over time. The *SID-meter* approach deals with a data stream as a potentially unbounded, implicitly ordered sequence of events $< e_1, e_2, ..., e_n, ... >$, such that each event is represented by an array of E measures.

SID-meter defines a sliding window to bound successive events to be considered to D_2 calculation. The window is divided into n_c time periods (named counting periods), each of which including a predetermined number of events (n_i). Therefore, $n_i \times n_c$ determines the size of the sliding window and n_i represents its movement step. The value of D_2 is then continually computed for the events inside the window and updated when new n_i events arrive. Figure 1 illustrates a three-dimensional data stream (attributes a_1, a_2, a_3) processed through a sliding window divided into five counting periods ($n_c = 5$).

3. THE PROPOSED APPROACH

This article presents a process to analyze multiple time series by combining data stream monitoring and time series clustering. Our approach deals with multiple time series as multidimensional data streams, i.e., each series is considered an attribute of the stream. For instance, temperature and precipitation time series are integrated to define a two-dimensional data stream.

The proposed analysis process has two main steps, as illustrated in Figure 2: 1) *off-line* processing of the data stream using an extension of the algorithm *SID-meter*; 2) clustering of fractal dimension measures computed in the previous step.

SID-meter was originally designed to consider only a single sliding window of a predetermined size. In order to be applied to climate data more appropriately, we extended the algorithm to support sliding windows of different sizes applied simultaneously through a single reading of the data stream. The windows are generated based on initialization parameters defined according to the interests of the

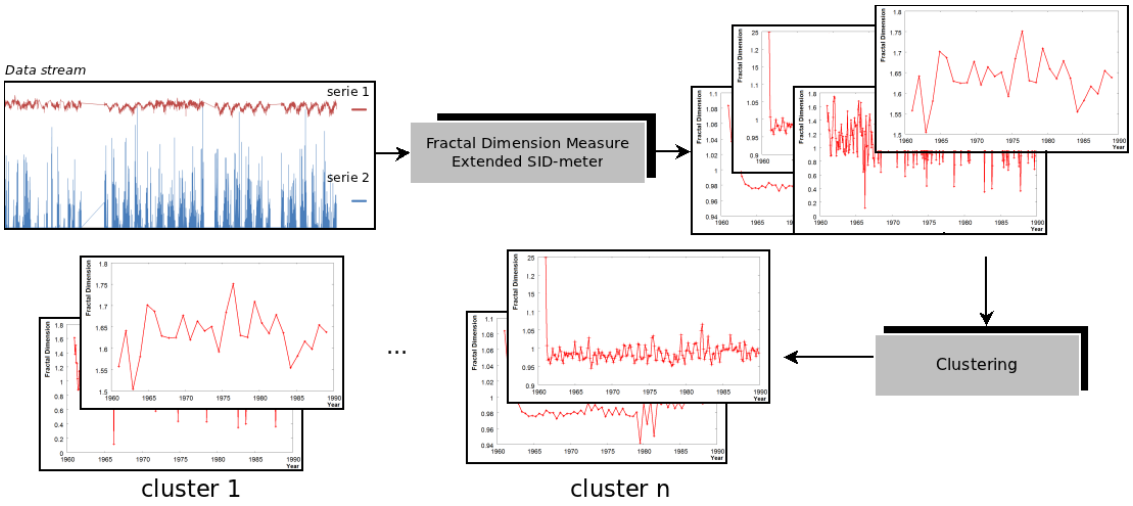

Fig. 2. Analysis process of multiple time series.

specialist: the smallest and the largest windows, i.e., minimum and maximum values of n_c (counting periods) and n_i (events per period). In addition, increment values for n_c and n_i are set to automatically create intermediate windows. As a result, *SID-meter* outputs several graphs indicating the variation of D_2 over time for different sliding windows. This extension allows temporal behavior analysis of the data stream in different granularities of time, aiming to detect patterns occurring monthly or annually, for instance.

Finally, the graphs generated by the *SID-meter* can be considered time series of D_2 measures. Thus, aiming at a detailed study, the D_2 series are clustered in order to identify similar patterns appearing in different granularities of time and patterns that are revealed only when the data are analyzed with more specific granularities. In this work, we used the *K-Medoids* partitioning method [Kaufman and Rousseeuw 1990] with the well-known *Dynamic Time Warping* (DTW) to measure the similarity between series. By performing a relaxation during the comparison of the patterns, DTW finds similarities even if there are displacements or deformations in the series. We implemented *K-Medoids* in our initial studies because it is simple, widely used and allows choosing elements of the dataset as the centers of the clusters.

4. EXPERIMENTAL RESULTS

We have applied the proposed analysis process to assess climate time series. Two datasets used in our experimental studies are detailed as follows:

1. Real data - climate time series provided by Agritempo[1] containing daily measurements of precipitation and mean temperature obtained from 25 ground-based meteorological stations of the state of São Paulo from 1961 to 1990.

2. Estimated data - climate time series composed of estimated measurements of mean temperature and precipitation obtained from *WMC Global Climate Resource* website[2]. These measurements are estimated by spatial interpolation using monthly climatological averages of real data gathered by meteorological stations, generating climate series for all points of a 0.5×0.5 global grid. We have selected estimated series from 1961 to 1990 considering grid points close to real meteorological stations of the Agritempo database (in São Paulo) in order to properly compare experimental results.

[1]Agrometeorological Monitoring System - http://www.agritempo.gov.br/
[2]http://climate.geog.udel.edu/∼climate

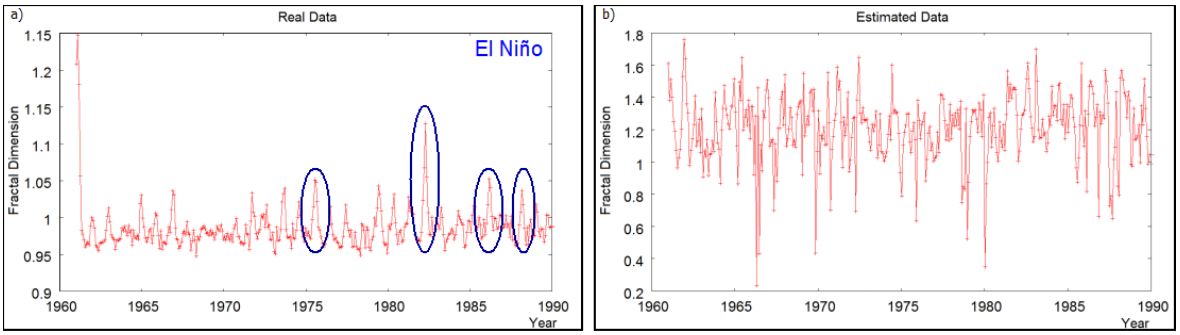

Fig. 3. Changes of D_2 for a three-month window: (a) real data - Agritempo; (b) data estimated by spatial interpolation - WMC.

A two-dimensional data stream composed of two attributes, `precipitation` and `mean temperature`, was defined for each dataset. We set the number of counting periods (n_c) varying from 2 to 5 and the number of events per period (n_i) from one month ($n_i = 30$) to one year ($n_i = 365$) as the initial parameters of *SID-meter*. It means windows ranging from two months to five years.

The graphs of fractal dimension variation (D_2) generated by *SID-meter* for both datasets showed significant differences in trends. Real data resulted in graphs with less variation of D_2, which remained around 1 as illustrated in Figure 3a. It indicates that the variables `precipitation` and `mean temperature` are correlated, as expected by the meteorologists. According to them, the correlation between these variables varies from a stronger correlation in some periods to a weaker correlation in others. On the other hand, graphs generated from data estimated by spatial interpolation showed more variation in the value of D_2 and abrupt changes over time. Moreover, values of D_2 are close to 2 indicating weak or no correlation between the variables, as it can be seen in Figure 3b. This result clearly indicates that methods usually applied in the Meteorology area to estimate measurements still require improvements, even from a purely numerical point of view.

The differences in the correlation between variables identified in real data and estimated data were also indicated through the execution of the *K-Medoids* clustering algorithm. Graphs (D_2 series) generated from real data and graphs generated from estimated data were respectively grouped into disjunct clusters. The clustering algorithm created six clusters of real data, each one including graphs related to windows of the same size but with distinct movement steps. On the other hand, considering only the estimated data, *K-Medoids* created four clusters with graphs related to different window sizes and movement steps. It is relevant to note that clustering the graphs related to different temporal granularities makes the analysis process easier for the experts, since window sizes can vary greatly. Thus, experts can select faster which time windows show the phenomenon being studied more clearly.

In a punctual analysis, the graph in Figure 3a shows the D_2 values for a three-month window with one month of movement step applied to real data. Although D_2 does not significantly exceed the value 1, the graph has some peaks that are highlighted in the figure. The biggest one matches the year 1983, when occurred a strong El Niño (1982/1983). The El Niño Southern Oscillation (ENSO), which is a large scale phenomenon that occurs in the Pacific Ocean (coast of South America), is characterized by the warming of surface waters in that region. ENSO usually causes intense rainfall in the Southern and droughts in the Northeast of Brazil [Berlato and Cybis 2003]. The state of São Paulo is specially influenced by the South Atlantic Convergence Zone (SACZ) according to studies on the influence of ENSO on rainfall regime in South America during monsoons [Grimm and Tedeschi 2009]. In El Niño years, the number of occurrences of intense rains increases from October to February with a break in January in the southeast region of Brazil. However, the signal is less pronounced considering the total of monthly rainfall occurring in some regions and not in others.

Other positive peaks in Figure 3a are also related to El Niño years, such as the peak in 1986 that

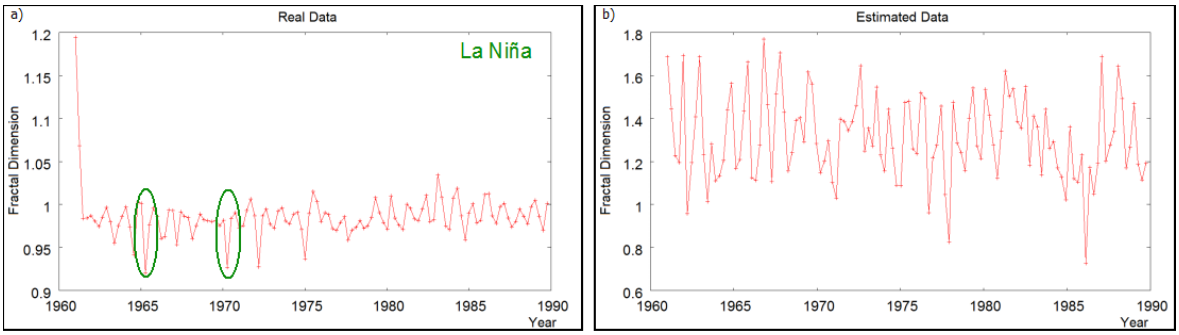

Fig. 4. Changes of D_2 for a six-month window: (a) real data - Agritempo; (b) data estimated by spatial interpolation - WMC.

corresponds to a moderate El Niño (1986/1988) and the peak in 1976 that refers to a weak El Niño (1976/1977). Thus, changes in the ratio of precipitation and temperature can be identified in the fractal dimension graphs. In other words, variations in the fractal dimension appear when trend changes occur in the variables being analyzed.

However, when analyzing graphs generated from estimated data, as illustrated in Figure 3b, we cannot identify a matching between fractal behavior and anomalies (climate phenomena). Probably, the generation process of climate series by spatial interpolation must have used few points of real measurements to estimate other points on the grid, causing several inaccurate estimates. In fact, it indicates that it is necessary to be careful and assume certain conditions to use interpolation methods in order to improve the coverage of areas with few meteorological stations. Thus, although spatial interpolation is a common practice employed by meteorologists, in specific conditions the interpolation results may even make the analysis process difficult to be accomplished due to distorting in correlations among different climate variables.

Data from the same period showed in Figure 3a were used to generate the graph presented in Figure 4a, but now using a six-month window with three months of movement step. Notice that this graph indicates the same general patterns as the previous one (Figure 3a) and therefore both of them were included in the same cluster by *K-Medoids*. As the sliding window is larger, the number of points in the graph of Figure 4a decreases and the positive peaks are more tenuous, as in 1983 which corresponds to an El Niño. On the other hand, negative peaks such as in 1966 and 1971 are coincident with occurrences of La Niña, which is an important anomaly characterized by cooling of surface water and increasing of atmospheric pressure in the eastern Pacific [Berlato and Cybis 2003]. In years of La Niña, the number of extreme events and their intensity are reduced. Therefore, there are stronger correlations between climate variables, as indicated by the values of D_2.

Finally, in contrast with real data, the graph generated from estimated data for the same window configuration (Figure 4b) shows no clear patterns in the D_2 behavior over time. Once more, this result indicates that estimates used to generate climate series do not accurately reflect the behavior of real data, and this is relevant information to meteorologists.

5. CONCLUSION AND FURTHER WORK

In this article we presented an approach to analyze multiple time series by combining data stream monitoring and clustering methods.

In general, initial results showed that the fractal-based analysis of climate time series can indicate behavior changes that coincide with climate phenomena. Furthermore, clustering of D_2 time series generated by *SID-meter* for different sliding windows showed that similar temporal patterns appear

in the same cluster. Thus, specialists can accomplish more refined studies considering temporal granularity in each cluster.

It is noteworthy that comparing series of real meteorological data with estimated data, as well as identifying when the estimated data show behavior actually compatible with real data, is fundamental to improve the meteorologists' work. In fact, research on improving the generation of estimated data is crucial due to the growing importance of more detailed studies on trends of climate scenarios, in particular when the number of meteorological stations is not sufficient for accurate studies.

Further work includes analysis of climate series from other regions of Brazil, which are also strongly influenced by climate phenomena, such as El Niño and La Niña. Moreover, we propose to assess other algorithms for clustering detection including specific algorithms for time series clustering. We also intend to perform comparative analysis considering other trend detection techniques.

ACKNOWLEDGMENT

Authors thank Embrapa, Fapesp, CNPq, Capes and Microsoft Research for the financial support and Agritempo for the climate data used in this work.

REFERENCES

Alexander, L., Zhang, X., Peterson, T., Caesar, J., Gleason, B., Tank, A., Haylock, M., Collins, D., Trewin, B., Rahimzadech, F., Tagipour, A., Kumar, K. R., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D., Burn, J., Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci, M., and Vasquez-Aguirre, J. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research* 111 (D05109): 1–22, 2006.

Ayoade, J. O. *Introdução à climatologia para os trópicos.* Ed.Brestrand Brasil, Rio de Janeiro, 1996.

Baioco, G. B., Traina, A. J. M., and Traina, C. Mamcost: Global and local estimates leading to robust cost estimation of similarity queries. In *Proceedings of the SSBBM International Conference on Scientific and Statistical Database Management.* Banff, Canada, pp. 6–16, 2007.

Barbará, D. and Chen, P. Using the fractal dimension to cluster datasets. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Boston, MA, USA, pp. 260–264, 2000.

Berlato, Moacir A., F. and Cybis, D. *El Niño e La Niña: impactos no clima, na vegetação e na agricultura do Rio Grande do Sul; aplicações de previsões climáticas na agricultura.* Editora da UFRGS, Porto Alegre, 2003.

Chakrabarti, D. and Faloutsos, C. F4: large-scale automated forecasting using fractals. In *Proceedings of the CIKM International Conference on Information and Knowledge Management.* McLean, VA, EUA, pp. 2–9, 2002.

Faloutsos, C. and Kamel, I. Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *Proceedings of the ACM PODS Symposium on Principles of Database Systems.* Minneapolis, MN, USA, pp. 4–13, 1994.

Grimm, A. and Tedeschi, R. G. Enso and extreme rainfall events in south america. *Journal of Climate* 22 (7): 1589–1609, 2009.

Kaufman, L. and Rousseeuw, P. J. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley and Sons, 1990.

Schroeder, M. *Fractals, Chaos, Power Laws.* W. H. Freeman and Company, 1991.

Sousa, E. P. M., Traina, C., Traina, A. J. M., and Faloutsos, C. Measuring evolving data streams' behavior through their intrinsic dimension. *New Generation Computing Journal* 25 (1): 33–59, 2007.

Sousa, E. P. M., Traina, C., Traina, A. J. M., Wu, L., and Faloutsos, C. A fast and effective method to find correlations among attributes in databases. *Data Mining and Knowledge Discovery* 14 (3): 367 – 407, 2007.

Traina, C., Sousa, E. P. M., and Traina, A. J. M. Using Fractals in Data Mining. In M. Kantardzic and J. Zurada (Eds.), *New Generation of Data Mining Applications.* Wiley/IEEE Press, pp. 599–630, 2005.

Traina, C., Traina, A. J. M., Wu, L., and Faloutsos, C. Fast feature selection using fractal dimension. In *Proceedings of Brazilian Symposium on Databases.* João Pessoa, PB, Brazil, pp. 158–171, 2000.

Zhai, P., Baethgen, W. E., Cerda, M. S., Davey, M., Goolaup, P., Kontongomde, H., Kousky, V. E., Llansó, P., Ropelewski, C. F., and Reid, P. Guidelines on climate watches. Tech. rep., World Meteorological Organization, 2005. http://www.wmo.int/pages/prog/wcp/wcdmp/documents/GuidelinesonClimateWatches.pdf.