Analysis of Large Scale Climate Data: How Well Climate Change Models and Data from Real Sensor Networks Agree?

Santiago A. Nunes
University of São Paulo - USP
São Carlos, Brazil
santiago@icmc.usp.br

Luciana A. S. Romani
Embrapa Agriculture Informatics
Campinas, Brazil
luciana.romani@embrapa.br

Ana M. H. Avila
State University of Campinas
Campinas, Brazil
avila@cpa.unicamp.br

Priscila P. Coltri
University of Campinas - Unicamp
Campinas, Brazil
pcoltri@cpa.unicamp.br

Caetano Traina Jr.
University of São Paulo - USP
São Carlos, Brazil
caetano@icmc.usp.br

Robson L. F. Cordeiro
University of São Paulo - USP
São Carlos, Brazil
robson@icmc.usp.br

Elaine P. M. de Sousa
University of São Paulo - USP
São Carlos, Brazil
parros@icmc.usp.br

Agma J. M. Traina
University of São Paulo - USP
São Carlos, Brazil
agma@icmc.usp.br

ABSTRACT

Research on global warming and climate changes has attracted a huge attention of the scientific community and of the media in general, mainly due to the social and economic impacts they pose over the entire planet. Climate change simulation models have been developed and improved to provide reliable data, which are employed to forecast effects of increasing emissions of greenhouse gases on a future global climate. The data generated by each model simulation amount to *Terabytes* of data, and demand fast and scalable methods to process them. In this context, we propose a new process of analysis aimed at discriminating between the temporal behavior of the data generated by climate models and the real climate observations gathered from ground-based meteorological station networks. Our approach combines fractal data analysis and the monitoring of real and model-generated data streams to detect deviations on the intrinsic correlation among the time series defined by different climate variables. Our measurements were made using series from a regional climate model and the corresponding real data from a network of sensors from meteorological stations existing in the analyzed region. The results show that our approach can correctly discriminate the data either as real or as simulated, even when statistical tests fail. Those results suggest that there is still room for improvement of the state-of-the-art climate change models, and that the fractal-based concepts may contribute for their improvement, besides being a fast, parallelizable, and scalable approach.

Categories and Subject Descriptors

H.2 [**Database Management**]: Database Applications

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

Keywords

Sensor Networks, Data Streams, Fractal Analysis, Climate Data

1. INTRODUCTION

A major challenge posed to researchers in the 21st Century refers to assess how much of the world climate changes as a consequence of the global warming are due to human activities. The unquestionable increase in the average temperature has impelled researches to do collaborative work involving meteorologists, mathematicians, statisticians and computer scientists, in order to assess the real impact of such increases and at what extent we have the ability to create strategies to deal with them as well. Provided that the global warming affects the entire planet, the Intergovernmental Panel on Climate Change (IPCC) [21] was created to evaluate and analyze such changes and to propose alternatives to deal with the current and future problems caused by climate changes. The IPCC reported that for the last hundred years, the average temperature on earth has been continually raising [19, 20, 32]. The ten warmest years registered in history are within the last twenty years [19]. 2010 is tied with 2005 as the warmest years on record [19, 25].

Climate changes forecast allows us to understand, and maybe to prevent and mitigate bad consequences of human activities. The forecasting task uses models derived from the iterative numerical solutions of differential equation systems, known as climate change models, which describe the main physical and dynamical processes of the climate system to simulate future climates as response to changes in the atmosphere and in the oceans [18, 14, 26]. Each execution of one climate model commonly takes weeks of processing in a large, highly-parallel processing computer and generates several *Terabytes* of time series data, depending on the number of climate parameters. Those parameters commonly go

up to a few tens and include temperature, humidity, wind direction and intensity, among others [28, 27, 4]. For the simulation, the atmosphere is divided into cubes spanning a few kilometers wide and hundreds of meters high, covering the whole region or the entire earth, up to the stratosphere. The models are usually evaluated starting the simulation at one given instant in the past, where the real conditions of the atmosphere were known, and using statistical comparison of trend analyses to compare the simulated results to the real data recorded thereafter from a network of sensors from meteorological stations existing in the analyzed region. Today, the analyses of statistical significance indicate that the results provided by state-of-the-art climate models closely follow the recorded data. Therefore, these statistical analyses give an evidence of the model's correctness. However, we have been performing additional analysis using Fractal Theory techniques, and we have found that those fractal-based techniques can clearly differentiate the simulated from the real data based on the intrinsic correlations among climate variables. Thus, although the current climate change models are appropriate from the statistical point of view, the Fractal Theory shows that they can be improved.

In such context, this paper proposes one novel process of analysis as an alternative strategy to evaluate the accuracy of climate change models when compared to real climate data, mainly considering general temporal behavior and correlations among climate variables. Our approach deals with multiple time series as one multidimensional data stream that corresponds to a set of real ground stations within the sensor network or to specific atmospheric cubes in the simulation, in a way that each time series (climate variable) defines an attribute of the stream. Therefore, it is possible to integrate multiple climate variables into one unified process of analysis, where the correlations existing among the variables can be analyzed. In particular, we continuously analyze the data using fractal-based data stream monitoring for change detection, considering the intrinsic correlation among time series defined by different climate variables. It is worth to note that our technique does not interfere in the simulation processing, so its accuracy is preserved.

Up to now, we performed experimental studies over the climate data series obtained from meteorological stations in a given region and the simulated corresponding data from a regional climate model for the same region. The results show that our approach can clearly discriminate the data either as real or as model generated. Those results suggest that, although the current climate change models are appropriate from the statistical point of view, there is still room for improvement of the models, and that the fractal-based concepts may play an important role in the task. Moreover, our process relies only on the subset of data collected from the ground stations of a given region and on the corresponding cubes in the simulated data. Therefore, our proposed evaluation can be performed over the network of weather monitoring centers, so each one can better improve its corresponding analyses.

It is important to highlight that the techniques for data analysis based on the Fractal Theory are specially well-suited for the analysis of very large collections of data. This is true mainly because of two facts: (i) the fractal-based approaches usually allow fast processing with linear or quasi-linear scalability regarding the number of data elements and attributes; and (ii) the fractal-based methods commonly rely on the partition of the data space, the individual analysis of each partition and the integration of the results, following the well-known "divide-and-conquer" strategy that clearly contributes to their parallelization. These characteristics exist in all techniques proposed in this paper. Thus, our techniques can be seamlessly parallelized to allow the analysis of data coming from global climate simulations as well as from worldwide networks of meteorological stations.

This paper is organized as follows: Section 2 presents background concepts and related work. Section 3 describes our approach to analyze climate time series coming either from networks of meteorological stations or from climate models. Experimental results are discussed in Section 4. Finally, Section 5 presents final remarks.

2. BACKGROUND AND RELATED WORK

This section presents background concepts of climate changes forecast and of the Fractal Theory applied to the analysis of data streams.

2.1 Climate Changes Forecast

Several research groups from different countries have been working with global climate models of similar characteristics [18, 14, 26]. The models accurately represent the atmospheric, oceanic and earth processes. Climate models are systems of differential equations derived from the basic laws of physics, fluid motion and chemistry. These equations are solved at a large number of points on a three-dimensional grid covering the entire world and, therefore, usually run on supercomputers. Note that these models are the only means to estimate the effects of increasing greenhouse gases on future global climate. It has been established that the amount of greenhouse gases in the atmosphere affects the amount of energy that escapes from Earth [17]. Observations in [17] have conclusively demonstrated that the abundance of greenhouse gases in the atmosphere has risen dramatically since the beginning of the Industrial Age, and that the amount of energy entering and escaping from Earth is the major determinant factor in climate. Therefore, changes in that balance - either in the input or in the output - will probably cause a directional change in climate [21].

Due to the inherently chaotic nature of the relationship between the atmosphere, earth and the oceans, it is not trivial to reach and guarantee accuracy of the models designed to forecast climate changes in the next years. Therefore, several future climate scenarios are possible, with different probabilities to occur, depending on the emphasis given to the input parameters settled for each model run. The uncertainty associated to the inherent changing of the climate behavior underlines the design of climate models. As distinct scenarios forecast different changes in the future, the consequences (e.g., in the agricultural production) can be large or small with different impacts for distinct regions. Thus, the better the accuracy of the model, the more dependable is the information provided to the decision makers. In this sense, the well-adjusted knowledge provided by the climate model may support strategic actions from governments and enterprises.

For example, studies were conducted aimed at verifying the impact that rising temperatures - besides the corresponding effects on water availability - can cause on the Brazilian agriculture by the end of the century [28, 27, 4]. The study was based on the climate model PRECIS [2, 23].

The results point out that the climate change will cause a migration of crops adapted to tropical areas to the south of the country or to areas of higher altitudes to compensate alterations on the climate conditions. At the same time, there will be a decrease in the fields of crops from temperate climate in the country. The only crop that can benefit from rising temperatures is the sugarcane.

Note, however, that global climate models operate with low-spatial-resolutions, and therefore, they do not provide accurate local information, which is needed, for instance, to help planning crop production. The most accepted approach to enhance the task of mapping the low-spatial-resolution from the global climate models to a finer resolution is combining downscaling from the forecasts generated by the global models with local climate models. These local models work with higher resolution data over a restricted region of interest, using global models as lateral boundary conditions [3]. Notice, however, that working with more than one model simultaneously contribute to increase the amount of data required and the complexity of handling it.

Since 1996 the regional model Eta-CPTEC is under development in the Brazilian Center for Weather Forecasts and Climate Studies (CPTEC) aimed at providing weather forecasts for South America [11]. The Eta model was initially developed at the University of Belgrade and the Hydrometeorologic Institute from former Yugoslavia, and afterwards by the National Centers for Environmental Prediction (NCEP) [8]. The model output is used to analyze the variability of changes at several scales, from daily to yearly cycles, considering 50 Km of regional spatial resolution. The training database contains the historical climate information from 1961-1990 provided by the Eta model, and it is parametrized using the conditions provided by the Hadley Centre. It is important to highlight that each execution of the Eta model generates among three to five *Terabytes* of data, depending on the number of parameters employed, which goes up to 42 attributes [28, 27, 4].

2.2 Fractal Theory Applied to the Analysis of Data Streams

In this work, we propose a fractal-based approach to both analyze the temporal behavior of data from climate models and to compare it with real climate data coming from networks of meteorological stations. In particular, we have used the fractal-based technique proposed in [30] to track the behavior of evolving data streams. This section summarizes background concepts from the Fractal Theory and their application on the analysis of data streams.

A fractal is characterized by the self-similarity property, i.e., it is an object that presents roughly the same characteristics when analyzed over a large range of scales. Therefore, parts of any size of a fractal present the same characteristics of the whole fractal [29]. Examples of real and synthetic fractals are shown in Figure 1.

From the Fractal Theory, the Correlation Fractal Dimension D_2 is particularly useful for data analysis, since it can be applied to estimate the intrinsic dimension of real datasets that exhibit fractal behavior, i.e., exactly or statistically self-similar datasets. Indeed, it is well-known in the Databases community that most real datasets are roughly self-similar [15, 34, 35]. The Correlation Fractal Dimension D_2 measures the non-uniform behavior of real data considering both linear and nonlinear attribute correlations [15, 33, 31]. Therefore, D_2 represents the dimensionality of the dataset regardless of the dimension E of the space defined by its attributes. For instance, a set of points defining a line $z = ax + by + c$ embedded in a three-dimensional space with dimensions $[X, Y, Z]$ (and thus $E = 3$) has $D_2 = 1$, as there is a linear correlation between its attributes. That is, if the set of points is in a two-, three- or any higher dimensional space, it will always keep its shape and organization (points along a line), and have the intrinsic dimension equal to one. See Figure 2 that illustrates this idea.

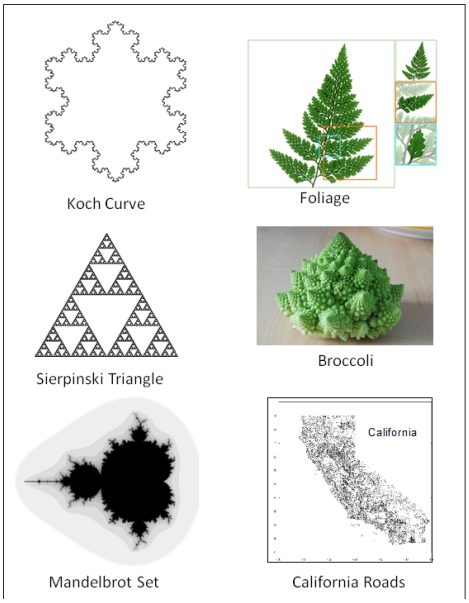

Figure 1: Examples of Fractals.

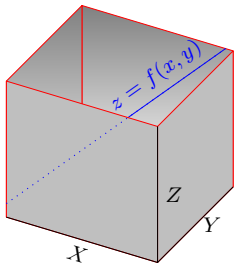

Figure 2: A line segment embedded in a 3-dimensional space has $D_2 = 1$.

The Correlation Fractal Dimension D_2 of E-dimensional real datasets can be computed by the *BoxCounting* method [29]. Equation 1 presents its definition, in which r is the side size of the cells in a (hyper) cubic grid that divides the address space of the dataset, $[r_1, r_2]$ is a significative range of scales, and $C_{r,i}$ is the count of points in the ith cell.

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

Concepts from the Fractal Theory have been applied to tackle several problems in databases and in data mining, such as selectivity estimation [16, 9, 5], clustering [7, 12, 13,

6], time series forecasting [10], correlation detection [31] and data stream analysis [30].

The work presented in [30] proposes a technique to detect changes in multidimensional, evolving data streams based on the information of intrinsic behavior provided by the fractal dimension D_2. The authors also present the algorithm *SID-meter* to continuously measure D_2 over time aimed at monitoring the evolving behavior of the data, such that significant variations in successive measures of D_2 can indicate changes in the intrinsic characteristics, as well as in attribute correlations in the data.

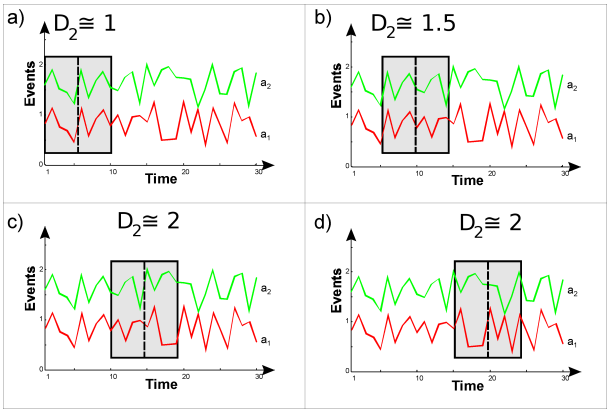

Figure 3: Sliding window over a bi-dimensional data stream.

The *SID-meter* deals with a data stream as a potentially unbounded, and implicitly sequence of events $< e_1, e_2, ... >$ ordered in time, such that each event is represented by an array of E measures, i.e., $e_i = (a_1, a_2, ..., a_E)$. To measure the fractal dimension of the stream over time, *SID-meter* uses a sliding window to bound successive events to be considered into the calculation of D_2. The sliding window is divided into n_c periods where each period is defined by a predetermined number of events n_i (or units of time), such that whenever n_i new events arrive, the n_i oldest events are discarded. In other words, $n_c * n_i$ specifies the length of the window and n_i determines the step by which the window moves. The size of the window and its movement step are user-defined parameters. The value of D_2 is continuously computed for the events inside the window and updated whenever new events arrive.

Figure 3 illustrates a bi-dimensional data stream (attributes a_1 and a_2) processed through a sliding window of size 10 units of time ($n_c = 2$ and $n_i = 5$). Notice in Figure 3a) that from time 1 to 10 (the first window), attributes a_1 and a_2 present a similar behavior. In fact, they are linearly correlated and therefore $D_2 \cong 1$. From Figure 3b) to 3d), old events are discarded while new data are processed, considering a movement step of 5 units of time. Also note that from time 10 on, the behavior of attributes a_1 and a_2 changes, and the attributes are no longer correlated, such that $D_2 \cong 2$. When the correlations between the attributes change, the value of D_2 changes as well. Therefore, by continuously measuring the fractal dimension, *SID-meter* outputs a sequence of D_2 values (see Figure 3) that monitors the evolving data behavior.

2.3 Methods for Evaluating Simulated and Real data

In order to evaluate how well simulated climate data behavior agrees with real observed data coming from networks of meteorological stations, some statistical methods are widely used, including the Pearson correlation coefficient (r), the coefficient of determination (r_2), the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and others [1]. However, those traditional measures are not always adequate to evaluate the agreement between real and simulated climate data, because of the inherent characteristics of the climate variables behavior. Usually, these methods indicate the amount of linear correlation among variable pairs and the single difference between data, which do not represent the true relationship between predicted and observed climate data. Moreover, they are also oversensitive to large extreme values (outliers) and insensitive to additive and proportional differences between model predictions and measured data [22], which could be normal in climate data behavior.

To circumvent these problems, Willmott [37, 39] developed the Index of Agreement d, which is used especially to validate models of prediction. It has been used satisfactorily when observed climate data and model-predicted data need to be compared. The advantage of this index is that it represents the ratio between the mean square error and the "potential error" [37, 39]. The Willmott index is defined by Equation 2:

$$d = 1 - \frac{\sum_{i=0}^{n}(P_i - O_i)^2}{\sum_{i=0}^{n}(|P_i - \bar{O}|) + (|O_i - \bar{O}|)} \qquad (2)$$

In Equation 2, P_i are the predicted values, O_i are the observed values and $\bar{O}$ is the mean of the observed values.

The numerator of the main term in Equation 2, $\sum_{n=1}^{\infty}|(P_i - O_i)|^2$, is the sum of the squared error (SSE). The denominator, which gives the sum of the squared absolute distances from P_i to $\bar{O}$ and from O_i to $\bar{O}$, is referred to as the potential error (PE). In fact, PE is dependent on the range of P and O, and it is used to standardized SSE. Thus, d is a bounded, non-dimensional measure which varies between 0 and 1. If all modeled values fit the observed values, as a perfect agreement, d equals 1, whereas a complete disagreement between predicted and observed values equals to zero. The attractiveness of Willmott's index is that it partitions the total error into systematic and unsystematic errors. The Willmott index of agreement can detect additive and proportional differences in the observed and simulated means and variances, which is well indicated to evaluate climate data predictions [24].

The Willmott index has been largely used by researchers because studies typically have different conclusions regarding the efficacy of the Pearson's correlation, tests of statistical significance (both parametric and non-parametric), and certain difference measures [1]. Those differences underscore the uncertainty that researchers face when testing a model, comparing two or more models, or selecting the most appropriate model from the literature [38, 36]. In this paper we take advantage of a fractal-based analysis together with the Willmott index d to evaluate how model-generated data and real climate data agree. This strategy allows us to identify linear, non-linear and even fractal attribute correlations by analyzing all climate attributes together, while the use of

other techniques (e.g., the Pearson's correlation) to perform the same task would identify linear correlations only, also demanding the analysis of every possible pair of attributes individually.

3. THE PROCESS OF ANALYSIS FOR REAL AND SIMULATED DATA

In order to assist the domain specialists (e.g., meteorologists) to study climate data, especially for the evaluation of the behavior of data from climate change models in comparison with that of real data recorded in meteorological stations from the network, we propose here a process of analysis based on the information provided by basic statistical measures and on the fractal dimension measurement, continuously calculated according to the *SID-meter* approach.

The process of analysis, illustrated in Figure 4, works as follows: time series defined by different climate variables are combined to create a multidimensional data stream, in a way that each variable defines a dimension of the stream. Following the *SID-meter* approach, each event e_i is defined as a set of climate measurements of different variables collected from networks of meteorological stations or generated by climate models, considering a specific location and a specific instant of time. An off-line implementation of the *SID-meter* algorithm processes the stream and computes the successive values of the fractal dimension over time, i.e., it outputs sequences of D_2 values. The size of the sliding window is defined by the specialist according to the target study. Hence, by analyzing the variation of the fractal dimension over time, it is possible to identify the existence of correlations among the climate variables and understand how such correlations evolve. In other words, it is possible to evaluate the temporal behavior of the data.

The fractal-based analysis can be applied to real climate time series and to time series generated by climate models (simulated data) as well, aimed at comparing them over time. The main idea is to evaluate the successive values of D_2 computed for both streams (simulated and real) focusing on the following aspects:

1. Analysis of the general behavior of the two sequences of D_2 values in order to evaluate significant variations of the fractal dimension, especially those changes occurring in periods characterized by well-known climate phenomena or extreme events.

2. Comparison of real and simulated data considering the previous analysis, aimed at evaluating similarities and discrepancies in the general behavior of the fractal dimension measured for both streams.

3. Comparison of the individual values of D_2 from both streams focusing on correlation analysis, i.e., comparing correlations among climate variables from real data and correlations among the same variables generated by the climate model.

In addition to the fractal analysis, statistical measures, such as mean and standard deviation, are computed for each climate variable (real and simulated), providing further information to assist the specialist in the analysis task, in particular when evaluating how the output of climate change models and data from real sensor networks agree.

We implemented the proposed process of analysis in the tool named *ClimFractal*, which was designed to support visual analysis of large climate datasets based on time series and stream mining algorithms. From the specialist point of view, the whole process is entirely performed through the *ClimFractal* interface. The specialist can pick the region of interest and visualize the geographical location of the sensors in ground-based meteorological stations of the network available for that region in the climate database. The climate time series gathered from the selected station define the data stream to be processed with *SID-meter*. The specialist also determines the time interval to be considered and defines the size and movement step of the sliding window according to the purposes of the analysis. Finally, simulated data from a climate model, related to the selected region and time interval, can also be included in the process as a second data stream and compared to the real data. The statistical measures and the successive values of the fractal dimension are computed for the time series and for the corresponding data streams of interest. Furthermore, the results are shown in multiple graphs to allow visual analysis of the temporal behavior of real data and simulated data from a climate model. Therefore, specialists can visually evaluate the results as well as tune the parameters, change the region and time interval of interest, and perform experimental studies with different configurations.

It is worthy to notice that one of the purposes of the aforementioned analysis, from the specialist perspective, is to evaluate climate models and sensor network data in the context of climate change researches, aiming at improving climate change forecasting. Also notice that our approach deals with very large sensor networks, because the *SID-meter* algorithm scales linearly with the size of the data generated by them, combining multidimensional time series into a sole data stream that summarizes their correlations regarding a time window. The ability of offering to the specialist a summarized means for spotting the correlation among attributes of the data is a valuable tool for real time data analysis.

4. EXPERIMENTAL RESULTS

We have applied the proposed process of analysis to assess climate time series from a climate database. Two sets of selected time series were used, which are described in Table 1 and detailed as follows.

1. `RealSeries` dataset: It has climate time series provided by Agritempo[1] with daily measurements of precipitation, and minimum and maximum temperatures obtained from a network of 283 ground-based meteorological stations in Brazil, from 1917 to 2010. Note that missing periods exist for some stations. This dataset spans $\sim$ 70 Megabytes and covers 283 stations. For the experiments reported, we used one part of this data with daily measurements of precipitation, and mean temperature obtained from 24 selected ground-based meteorological stations (those without considerable missing periods) of the state of São Paulo, Brazil, from 1961 to 1990, in order to avoid having to deal with missing data.

[1] http://www.agritempo.gov.br

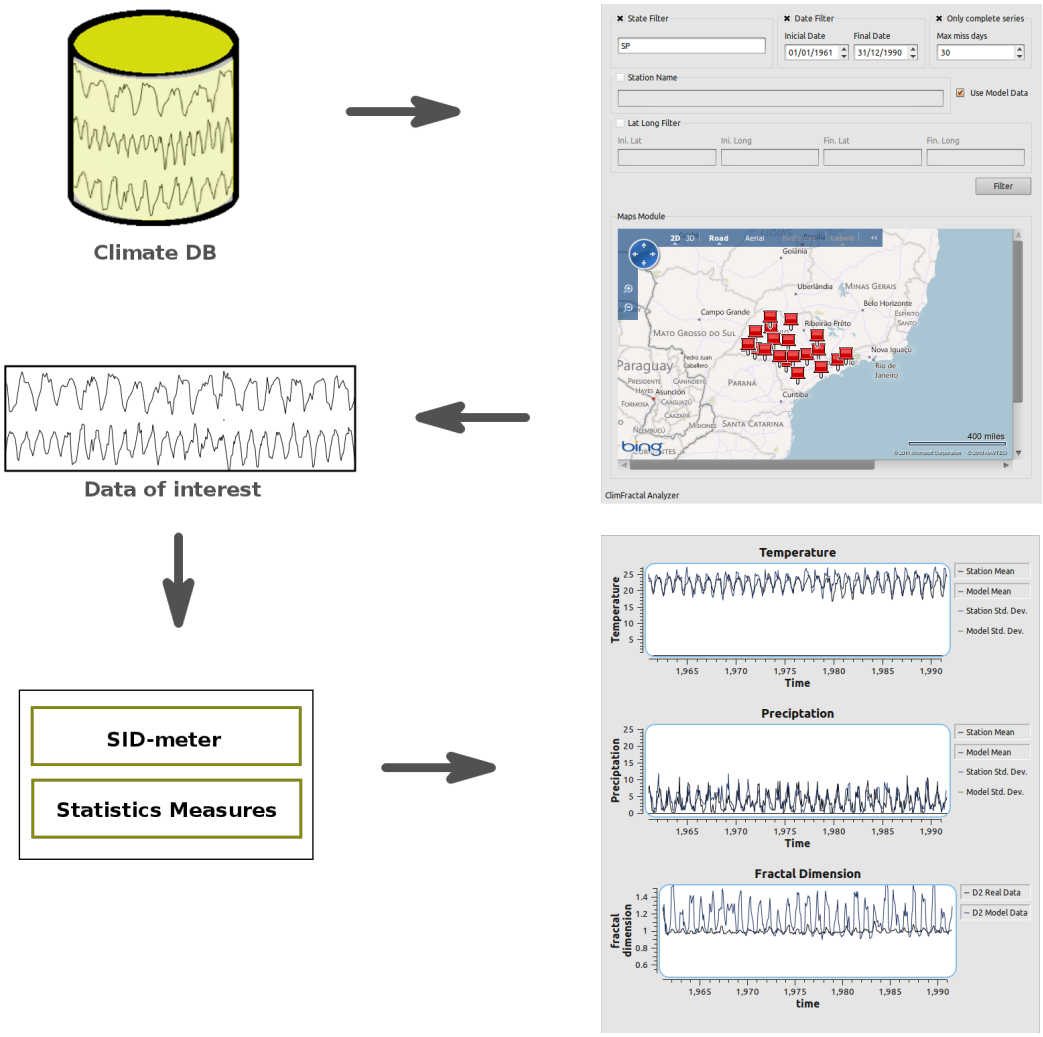

Figure 4: Overview of the proposed process of analysis.

2. `SimSeries` dataset: It has climate time series with estimated measurements for 42 atmospheric measures, such as precipitation, temperature, humidity and pressure, using the Eta-CPTEC climate model. Specifically, the data refers to four distinct simulations for all grid points on a 0.4×0.4 (Latitude $\times$ Longitude) grid that covers the entire South America and a small portion of Central America, with four values per day for each atmospheric measure; the simulation time ranges are: 1960 – 1990 and 2010 – 2099. The four simulations used different "what-if" scenarios, for the amount of greenhouse gases in the atmosphere expected in the future, ranging from a pessimistic one with high gas emissions, to an optimistic one with low gas emissions. Each of the four simulation scenarios generates ~ 1.25 *Terabytes*, with a total of ~ 5 *Terabytes* for the entire `SimSeries` dataset. In our experiments, we used the simulated values that correspond to the real data of the `RealSeries` dataset taken from the standard scenario, i.e., daily measurements of precipitation and mean temperature obtained from that scenario for the years between 1961 to 1990, considering points of the grid that are the closest to each real meteorological station of the `RealSeries` dataset. The results obtained highlights that the process can indeed be distributed so each region can adapt the analysis process results to target its specific requirements.

A two-dimensional data stream composed of the attributes precipitation and mean temperature was defined for each dataset. We performed the experiments with two configurations of sliding windows:

1. A three-month window evaluating the fractal dimension at each month, and;

2. A six-month window evaluating the fractal dimension at each two months.

The successive values of the fractal dimension computed by the *SID-meter* method for both streams as well as the statistical measurements are visualized in the *ClimFractal* interface (Figures 5 and 6). The measurements for the real data are presented in black, while those for the simulated data are presented in blue. It can be observed that both

Table 1: Summary of datasets.

Dataset	Number of Climate Variables	File Size
`RealSeries`	3	$\sim$ 70 Megabytes
`SimSeries`	42	$\sim$ 5 *Terabytes*

datasets showed a similar pattern of behavior, i.e., the fractal dimension measurements vary on the real data as well as on the Eta-CPTEC model time series.

However, it is important to notice that the real data resulted in graphs where the variables are more correlated, that is, with less variation of D_2. In fact, we can see that the correlation fractal dimension remained always around 1.0, as shown in Figure 5. This result indicates that the variables precipitation and mean temperature are highly correlated, as expected by the meteorologists. In fact, it is well-known in meteorology that the correlation between these variables varies from a stronger correlation in some periods to a weaker correlation in others, depending on the season of the year.

On the other hand, the values of D_2 are higher for the data estimated by the Eta-CPTEC model, with D_2 varying from 1 to 1.5, as it can be seen in Figure 6. Those values indicate that the correlation between the variables simulated in the climate model is smaller than that found in the real data. Even though, the similar patterns observed on both curves indicate that the model can represent the general climate in the period between 1960 to 1990 for the area of study, but it misses the finer details of the intrinsic characteristics of the climate system. Thus, we can see that fractal-based measurements reveal the expected correlations on the real data, and that they are not echoed when the same measurements are applied over simulated data. Note that the Willmott index and other statistical-based methods do not spot neither of them. These results reinforce the fact that the Willmott index d is not a measure of correlation, but rather it is a measure of the degree to which a model's predictions are free of errors. Willmott index normally is used, by specialist, to measure the model performance. In this case, if the index is hight we can say that the model is very closer and can reproduce the real data.

We found the same pattern of correlation between the climate variables using different window sizes. Although the pattern is similar for both windows: 3 months and 6 months (from Figures 5 and 6, respectively), the similarity between the fractal dimension graphs and their global behavior are more evident in windows of 3-months. This fact reinforces the need of specialized tools for visual analysis, in which experts can interact with the results by varying the input parameters to verify the response of the model for different periods of time.

The similarity pattern indicated in the fractal dimension graphs generated using real data and the output of the Eta-CPTEC's model was also observed when the conventional method of Willmott was applied over both datasets. The index of agreement d was computed for three climate variables (minimum temperature, maximum temperature and precipitation) considering data collected from a subset of meteorological stations in the network and the corresponding model-generated data. The index d presented a suitable

Table 2: Willmott measure for some stations at the State of São Paulo for 3 climates variables

Station Name	Min. Temp.	Max. Temp.	Prec.
AVARE	0.760473	0.727417	0.614577
PIRACICABA	0.716948	0.71237	0.676381
ALBERTO L.	0.824454	0.583125	0.534299
BOTUCATU	0.719529	0.706638	0.641365
SAO MIGUEL	0.850551	0.630149	0.563136

similarity pattern considering real and Eta-CPTEC's data, as it can be seen in Table 2. Most values of d for minimum temperature, maximum temperature and precipitation are greater than 0.6, indicating that the real and the simulated data follow one similar behavior, considering the individual behavior of each climate variable.

The proper agreement between real and Eta-CPTEC's data can be observed, for example, in the attribute of minimum temperature. On the other hand, the series containing precipitation measurements presented unsatisfactory results.

Values of d for minimum temperature were always greater than 0.72, indicating a satisfactory value of correlation. The station of SAO MIGUEL (Brazilian city of São Miguel Arcanjo) presented the best value of d (0.85). However, this region presented average values for d considering precipitation (0.56) and maximum temperature (0.63). The best value for d regarding maximum temperature (0.727) occurred in the analysis of data from the station of AVARE (city of Avaré), while average values of d to precipitation were reached for the station of PIRACICABA (city of Piracicaba). Note that these regions are located in an important area of sugarcane production. Thus, it is important to know the corresponding local and regional climatic characteristics in order to support the government on decision making analysis regarding planing of production. That conclusion highlights the importance of presenting to the analysts of each center in the weather monitoring network the ability to interpret the results from their own perspectives and requirements.

In addition, these results can also support the researchers that would use the Eta's model data, because they can evaluate the simulated data, indicating the strongest and the weakest correlations. The knowledge of climate variables and of their behavior recorded at networks of meteorological stations and estimated by climate models allow one to specify the tendency of local agricultural production both in the current climate as well as in adverse weather conditions in the future. It also provides the developers of the model feedback on how to better follow the real weather conditions in future releases of the model, integrating the overall behavior of the simulated data.

5. CONCLUSIONS

In this paper we proposed a process of analysis based on fractal concepts to evaluate climate data aimed at improving climate change research. It is important to highlight that having an approach that can summarize a large volume of multidimensional data (time series) generated by climate models as well as large networks of sensors, spotting their correlations is a valuable asset to support decision making processes. To do that, a fractal-based approach developed to monitor evolving data streams is applied to compare the

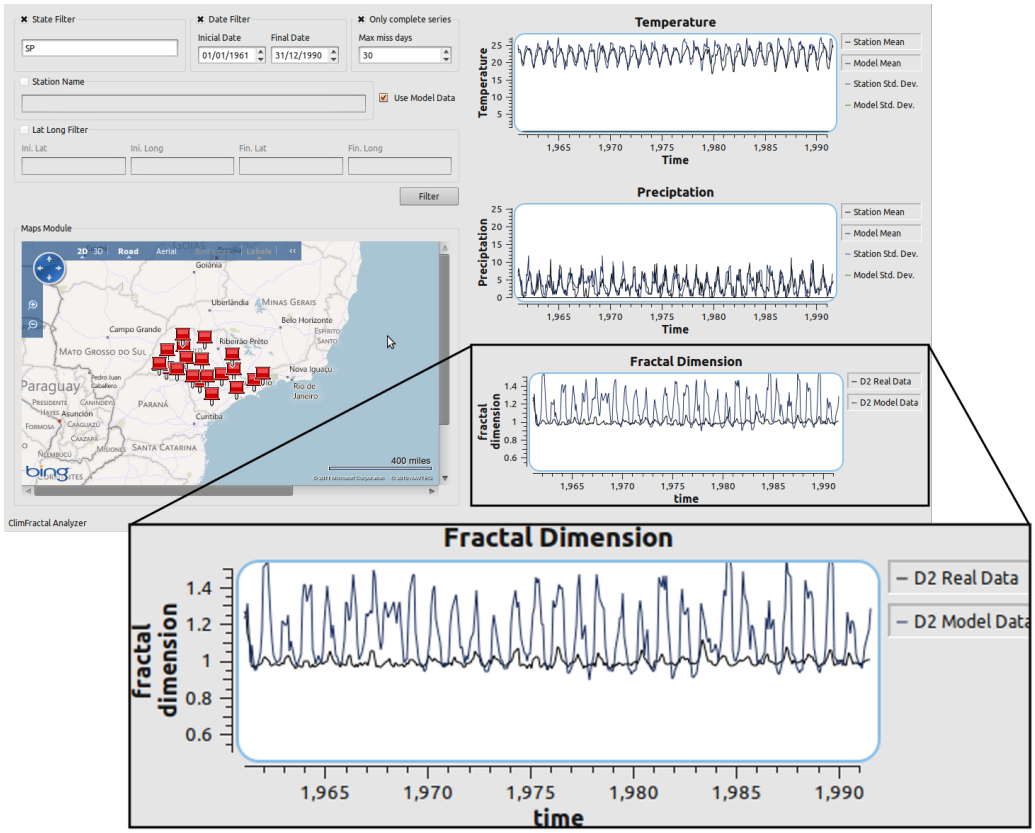

Figure 5: *ClimFractal* Interface: Fractal dimension measured with a 3-month window updated every month.

behavior of real climate time series with data from a climate model. Our proposed approach scales linearly on the dataset size and can be tuned to follow the seasonability of the data, what is done by adjusting its window size. We took advantage of a fractal-based analysis together with the Willmott index to evaluate the accuracy of climate models, mainly with respect to the correlations among climate variables. This strategy allowed us to identify linear, non-linear and even fractal attribute correlations by analyzing all climate attributes together, while the use of other techniques (e.g., the Pearson's correlation) to perform the same task would identify linear correlations only, also demanding the analysis of every possible pair of attributes individually. The whole process of analysis was implemented in the tool named *ClimFractal* System to support visual analysis of the results. It allows either the analysts of each center in the weather monitoring network to interpret the results from their own perspectives and requirements, or regulatory agencies spanning entire countries or continents to forecast future weather tendencies. This tool may be used by agro-meteorologists working with the production of intelligence data for government agencies that regulate funding of farms, aimed at stimulating the development of agricultural activities related to products that can admittedly provide better productivity in each region or in entire continents.

The initial results showed that our approach can discriminate between the real data coming from a network of sensors from meteorological stations and the data generated by a climate model, as the intrinsic correlations between climate variables identified in real data (and confirmed by the specialists) are considerably different from those generated by the climate model. These results suggest that there is yet room for improvement of the climate change models, and that the fractal-based concepts may contribute to this task.

The analysis of the output of climate change models can help specialists to better understand and to improve the model, thus contributing to the research on climate changes and their effects, such as in scenarios of positive and negative impacts on agriculture and production, as well as to the human being wellness.

Finally, it is important to highlight that the techniques for data analysis based on the Fractal Theory are specially well-suited for the analysis of very large collections of data. This is true mainly because of two facts: (i) the fractal-based approaches usually allow fast processing with linear or quasi-linear scalability regarding the number of data elements and attributes; and (ii) the fractal-based methods commonly rely on the partition of the data space, the individual analysis of each partition and the integration of the results, following the well-known "divide-and-conquer" strategy that clearly contributes to their parallelization. These characteristics exist in all techniques proposed in this paper. Thus, our techniques can be seamlessly parallelized to allow the analysis of data coming from global climate simulations as well as from worldwide networks of meteorological stations.

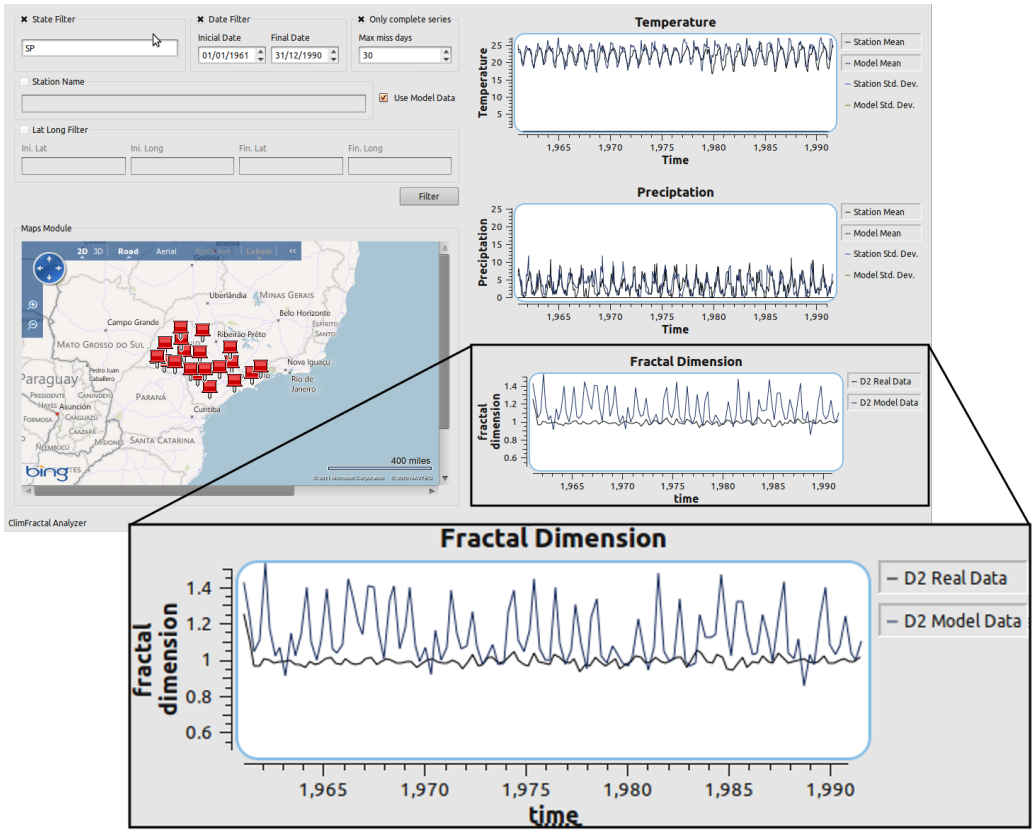

Figure 6: *ClimFractal* Interface: Fractal dimension measured with a 6-month window updated every 2 months.

Acknowledgments

The authors thank Embrapa, FAPESP, Microsoft Research, CNPq and Capes for the financial support, and Agritempo and CPTEC/INPE for the climate data used in this work. We also thank Dr. Chou Sin Chan from CPTEC/INPE for providing valuable support to the development of this work.

6. REFERENCES

[1] P. Ahlgren, B. Jarneving, and R. Rousseau. Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6):550–560, 2003.

[2] L. M. Alves and J. A. Marengo. Assessment of regional seasonal predictability using the PRECIS regional climate modeling system over south america. *Theoretical and Applied Climatology*, 100:337–350, 2010.

[3] T. e. a. Ambrizzi. Cenários regionalizados de clima no brasil para o século xxi: projeções de clima usando três modelos regionais: relatório 3. Technical report, MMA, Brasilia, 2007.

[4] E. D. Assad, H. S. Pinto, and J. J. Zullo. Impacts of global warming in the brazilian agroclimatic risk zoning. In *A Contribution to Understanding the Regional Impacts of Global Change in South America*, pages 175–182, São Paulo, Brazil, 2007. Instituto de Estudos Avançados da USP.

[5] G. B. Baioco, A. J. M. Traina, and C. Traina. Mamcost: Global and local estimates leading to robust cost estimation of similarity queries. In *SSDBM 2007*, pages 6–16, Banff, Canada, 2007. ACM Press.

[6] D. Barbará and P. Chen. Fractal mining - self similarity-based clustering and its applications. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 573–589. Springer, 2010.

[7] D. Barbará and P. Chen. Using the fractal dimension to cluster datasets. In *ACM SIGKDD*, pages 260–264, Boston, MA, 2000.

[8] T. Black. The new nmc mesoscale eta/cptec model: Description and forecast examples. *Forecasting*, 9:265–278, 1994.

[9] C. Böhm. A cost model for query processing in high dimensional data spaces. *ACM TODS*, 25(2):129–178, 2000.

[10] D. Chakrabarti and C. Faloutsos. F4: large-scale automated forecasting using fractals. In *CIKM*, volume 1, pages 2–9, McLean, VA - EUA, 2002. ACM Press.

[11] S. C. Chou, J. A. Marengo, A. A. Lyra, G. Sueiro, J. F. Pesquero, L. M. Alves, G. Kay, R. Betts, D. J. G. Chagas, L. Jorge, Bustamante, and P. Tavares. Downscaling of south america present climate driven by 4-member hadcm3 runs. *Springer - ClimDyn*, 25:33–59, 2007.

[12] R. L. F. Cordeiro, A. J. M. Traina, C. Faloutsos, and C. Traina. Finding clusters in subspaces of very large, multi-dimensional datasets. In *Proceedings of the 26th International Conference on Data Engineering (ICDE 2010)*, pages 625–636, Long Beach, California, USA, 2010. IEEE.

[13] R. L. F. Cordeiro, A. J. M. Traina, C. Faloutsos, and C. Traina. Halite: Fast and scalable multiresolution local-correlation clustering. *IEEE Trans. Knowl. Data Eng.*, 25(2):387–401, 2013.

[14] Djuric and Dusan. *Weather Analysis - Chapter I*. Prentice-Hall Inc., 1994.

[15] C. Faloutsos and I. Kamel. Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *ACM PODS*, pages 4–13, Minneapolis, MN, 1994.

[16] C. Faloutsos, B. Seeger, A. J. M. Traina, and C. Traina. Spatial join selectivity using power laws. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00)*, pages 177–188, Dallas, USA, May 2000.

[17] C. B. Field, V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G. K. Plattner, S. K. Allen, M. Tignor, and P. M. Midgley, editors. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2012.

[18] P. Forster, V. Ramaswamy, P. Artaxo, T. Berntsen, R. Betts, D. W. Fahey, J. Haywood, J. Lean, D. C. Lowe, G. Myhre, J. Nganga, R. Prinn, G. Raga, M. Schulz, and R. V. Dorland. *2007: Changes in Atmospheric Constituents and in Radiative Forcing*. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, 2007.

[19] Intergovernmental Panel on Climate Change – IPCC. *Climate change 2007: Fourth assessment report (AR4)*. Cambridge University Press, Cambridge, UK, 2007.

[20] Intergovernmental Panel on Climate Change – IPCC. *Climate Change 2007: Summary for Policymakers*. Cambridge Univ. Press., 2007. Formally agreed statement of the IPCC concerning key findings and uncertainties contained in the Working Group contributions to the Fourth Assessment Report.

[21] IPCC. Intergovernmental panel on climate change. http://www.ipcc.ch/ipccreports/index.htm, 2007. accessed: March, 2009.

[22] D. R. Legates and G. J. McCabe. Evaluating the use of ''goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Res.*, 35(1):233–241, 1999.

[23] J. A. Marengo, R. Jones, L. M. Alves, and M. C. Valverde. Future change of temperature and precipitation extremes in south america as derived from the PRECIS regional climate modeling system. *International Journal of Climatology*, 29(15):2241–2255, 2009.

[24] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3):885–900, 2007.

[25] NASA's Goddard Institute for Space Studies. Nasa research finds 2010 tied for warmest year on record. [Online]. Available at: `http://www.nasa.gov/topics/earth/features/2010-warmest-year.html`, Day of access: September 1, 2012.

[26] Petersen. *Weather Analysis and Forecasting*. 1956.

[27] H. S. Pinto and E. D. Assad. Global warming and the new geography of agricultural production in Brazil. page 42pp, Brasília, Brazil, 2008. The British Embassy.

[28] H. S. Pinto and E. D. Assad. Impacts of climate change on brazilian agriculture. In *Brazil: Assessment of the Vulnerability and Impacts of Climate Change on Brazilian Agriculture*. Development report for World Bank Project P118037, 2012.

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

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

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

[32] The National Academies. *Understanding and Responding to Climate Change: Highlights of National Academies Reports*. The National Academies, 2008.

[33] C. Traina, E. P. M. Sousa, and A. J. M. Traina. Using fractals in data mining. In M. M. Kantardzic and J. Zurada, editors, *New Generation of Data Mining Applications*, volume 1, pages 599–630 (Chapter 24). Wiley/IEEE Press, 2005.

[34] C. Traina, A. J. M. Traina, L. Wu, and C. Faloutsos. Fast feature selection using fractal dimension. *Journal of Information and Data Management - JIDM*, 1(1):3–16, 2010.

[35] C. Traina, A. J. M. Traina, L. Wu, and C. Faloutsos. Fast feature selection using fractal dimension - ten years later. *Journal of Information and Data Management - JIDM*, 1(1):17–20, 2010.

[36] C. J. Willmott. On the validation of models. *Physical Geography*, 2:184–194, 1981.

[37] C. J. Willmott. *On the evaluation of model performance in physical geography*. Gaile and Willmott, eds. Norwell, 1984.

[38] C. J. Willmott, R. Davis, J. Feddema, K. Klink, D. Legates, C. Rowe, S. Ackleson, and J. O'Donnell. Statistics for the evaluation and comparison of models. *JOURNAL OF GEOPHYSICAL RESEARCH*, 90:8995–9005, Sept. 1985.

[39] C. J. Willmott, S. M. Robeson, and K. Matsuura. A refined index of model performance. *International Journal of Climatology*, 32(13):2088–2094, 2012.