EL NINO TELECONNECTIONS RESEARCH: INITIAL RESULTS USING A MACHINE LEARNING AND DISCOVERY APPROACH

by

Z. Li
M. Kafatos
R. S. Michalski

Reports of the Machine Learning and Inference Laboratory , MLI 97-13,, George Mason
University, Fairfax, VA, 1997.

El Nino Teleconnections Research: Initial Results Using a Machine Learning and Discovery Approach

Z. Li, M. Kafatos, and R.S. Michalski

P 97-18
MLI 97-12

El Nino Teleconnections Research: Initial Results Using A Machine Learning and Discovery Approach

Zuotao Li & Menas Kafatos

Center of Earth Observing and Space Research (CEOSR)

Ryszard S. Michalski*

Machine Learning and Inference Laboratory

George Mason University, Fairfax, VA 22030

*Also with the Institute of Computer Science,

Polish Academy of Sciences, Warsaw, Poland

Abstract

The El Nino Southern Oscillation (ENSO) is the largest known global climate phenomenon on inter-annual time scales. It can have devastating social-economic consequences and is responsible for a large number of weather-related disasters. Thanks to the advent of modern remote sensing technology, understanding of ENSO has progressed substantically in the last decade. However, the relationship between El Nino signals and other geophysical parameters, especially those that may indicate teleconnections still remains unknown. In this work, the inductive learning system AQ15c, has been applied to explor the geophysical global satellite datasets in order to investigate possible teleconnections between the El Nino signals and other climate parameters. Results of experiments produce testable hypotheses and provide insights in understanding the El Nino teleconnection phenomenon.

Acknowledgments

We acknowledge useful inputs from Zhaoming Zhang, Xiangdong Ma, Lei Sun, and Dexiamg Liao.

The authors also acknowledge ideas and support from the Virtual Domain Application Data center (VDADC), Center for Earth Observing and Space Research (CEOSR) at George Mason University. Partial funding support from the NASA ESDIS Project (NAG 5-3086) and from the Goddard Global Change Data Center (NCC 5-143) is acknowledged.

The authors also thank for Machine Learning and Inference Laboratory (MLI) at George Mason University for providing software and assistance in using the AQ15c system in this project. The Laboratory's activities are supported in part by the Defense Advanced Research Projects Agency under Grant No. F49620-95-1-0462 administered by the Air Force Office of Scientific Research, in part by the National Science Foundation under Grants No. DMI-9496192 and IRI-9020266, and in part by the Office of Naval Research under Grant No. N00014-91-J-1351.

The authors wish to thank the Distributed Active Archive Center at Goddard Space Flight Center, Greenbelt, MD, 20771, for producing the data in its present format and providing them to the authors. The original data products were produced by the NOAA/NESDIS/NCDC Climate Analysis Center, Ashville, NC. Goddard's share in these activities was sponsored by NASA's Mission to Planet Earth program.

1. Introduction

The El Nino/Southern Oscillation (ENSO) is the largest known global climate signal on interannual time scales. According to Vallis (1988) who has attempted a definition of an El Nino event:

(1) El Nino occurs when an anomalously warm pool of water in the eastern equatorial Pacific Ocean is formed around late spring to begining of summer for southern latitutes. This event lasts for several months.

(2) Concurrent with the ocean warming, an atmospheric event occurs, namely a notable weakening of the trade winds. As one indicator, the sea level pressure difference between Darwin (Australia) and Tahiti is correlated with the oceanic events. This difference, termed the Southern Oscillation Index (SOI) measures an oscillatory event known as the Southern Oscillation (SO). El Nino events depart from the regular annual SO events.

(3) The ENSO events occur aperiodically, with intervals of between 2 and 11 years, nine or so events since 1945, when reliable records began. There is evidence for El Nino events for at least 400 years (Quinn et al., 1987). The strongest recent ENSO was the unusual 1982 event. An unusually strong warming event is currently developing in the tropical Pacific.

(4) The ENSO event is phased-locked to the seasonal cycle, normally reaching its maximum amplitude around Christmas time, althrough ENSO's usually start up to 6 months before and may last 1-2 years.

(5) Similar events, if they occur at all, are much weaker in the Atlantic and Indian Oceans.

(6) The opposite event, which is a cool event, is termed La Nina.

ENSOs are responsible for a large number of weather-related disasters, dramatic changes in marine ecology in tropical islands and coastal regions, and severe economic losses due to flood, famine and crop failure in countries along the tropical and subtropical belt, most notably in Africa. The effects of ENSOs are felt as far north as the United States where floodings and droughts may occur at different parts of the country.

The El Nino phenomenon is a very complex phenomenon and is thus difficult to provide with a simple explanation (Deser et al, 1992). The El Nino event occurs when the west-east gradient of surface water temperature weakens. Then the usual easterlies slacken and the height of water decreases. The pressure gradient between the eastern south Pacific and Indonesia is also weakened. The weakened surface circulation causes a large reduction in the rise of cold deep water in the east. The lowering of the height of water in the west may be accompanied by the induction of equatorial Kelvin waves which propagate to the eastern extremity of the Pacific, near the coast of South America off Peru, taking about 2-3 months to arrive there from the west, and contributing to the warming of surface waters. The change of temperature and the heavy rainfall which are both associated with the event can cause large ecological disasters.

Early researches had provided clues to understand this phenomenon, a fully interactive and coupled system which in the real world may be exhibiting characteristics of a 'chaotic' system, in that switches of state may be triggered by very small amplitude fluctuations or noise (Harries, 1994). Nervertheless, a chaotic system can exhibit quasi-stable periodicity, and even some degree of predictability.

Many weather and climate anomalies in the regions beyond major El Nino impacted areas, such as America, China, etc., are believed to have links with the El Nino signals and lead to the notion of "teleconnection". These teleconnections are associated with higher frequency, 30-60 day oscillations which manifest as oscillations in the upper tropospheric wind field. Understanding teleconnections could benefit a specific region by predicting and preventing natural disasters, and also have a great significance in the global climate change research. Climate anomalies provide opportunities as nature's "experiments" for researchers to examine various hypotheses on the impacts of Earth's changing climate. Thanks to the advent of modern satellite and remote sensing technology, global patterns depicted by the satellite observations from space have made this study possible.

In this work, the AQ15c system (Michalski, et al, 1986), an inductive learning system, has been applied to mine the geophysical global satellite datasets to investigate possible teleconnections between El Nino signals and other climate parameters. A set of parameters were chosen to perform the data mining experiment; data prepocessing & transformation algorithms are developed in the Virtual Domain Application Data Center prototype (VDADC) (Kafatos et al, 1997). Some interesting results are presented and discussed in this paper.

2. Method

Satellite observations from space provide scientists a good means to help solve the puzzle of El Nino. A major challenge facing Earth scientists today is the unavailability of high-level data analysis tools with which to extract useful information from the massive data sets produced by existing or future satellite observations. The data mining process in this project is interactive and iterative, involving numerous steps. It is based on the application of the inductive learning program AQ15c (Michalski et al., 1986; Wnek et al, 1995).

(1) Determining the application goal

In order to identify the El Nino signals, the features of the signals need to be extracted from the data based on the chosen attributes. This becomes a classification problem to distinguish El-Nino events from non El-Nino events. Since the El Nino is a very entangled, complex, and global scale natural phenomenon, the attempt of using the obtained classification rules to give a reasonably correct prediction of El Nino events is impossible without involving a complex climate simulation model. Thus, the application goal is not to use the classification rules to predicate the El Nino events, but to examine the classification rules themselves and attempt to find clues for the teleconnections between El Nino signals and other climate parameters.

(2) Selecting datasets for the experiment

A set of geophysical parameters (attributes) are chosen based on the domain knowledge and Datasets are selected based on the chosen geophysical parameters. One of the most important features of scientific data mining applications is the ability to use prior domain knowledge during the mining process. Actually, some researchers believe that data mining application in science may generally be easier than applications in other areas, such as business and finance, because science users typically know their data in detail and the background domain knowledge is usually available (Fayyad et al, 1996). For example, scientists could determine related attributes (parameters) and unrelated attributes (parameters) based on the domain knowledge. For this project several related attributes and their corresponding datasets are chosen.

(3) Data preprocessing and transformation

All the chosen geophysical datasets have been preprocessed for noise removal and calibration (see the dataset description part for details).

Based on the application goal, two steps of data transformation are performed in this work:

Step 1: Calculate the average parameter values in the large scale spatial regions which were previously selected based on the science research interests. One justification of doing this is that the El Nino phenomenon is a large spatial scale phenomenon. Thus, the detailed variations at smaller spatial scales are not of much interest.

Step 2: Discretize the average values for various parameters into numerous predefined levels where a uniform distribution is assumed (see appendix for the details).

After the two steps of data transformations, the target dataset volumes are reduced dramatically. (Some information may be lost due to the data transformations, but again, we are only interested in large scale variation trends in this application. A brief discussion can be found in the discussion part).

(4) Data integration and missing attribute value handling

Data from all target datasets are integrated to form one dataset for the mining process. The strategy of handling missing attribute values in this project is simply discarding the records which have missing data fields.

(5) Choice of the data mining algorithm (tool)

Based on the application goal, an appropriate data mining algorithm (tool) should be chosen. For the scientific applications, normally there are two requirements needed to be emphasized:

a. The discovered pattern (the form of knowledge representation) should be relatively easy to interpret by humans. This is very crucial for scientific research.

b. The data mining system should be able to benefit from the domain knowledge. One advantage of scientific application is that the scientific domain knowledge is well-documented and can be easily retrieved. How easy the domain knowledge can be incorporated into the data mining system also depends on the form of knowledge representation.

Based on these two requirements. AQ15c was chosen as our data mining tool. AQ15c system is an inductive learning system that generates decision rules. Rules are represented in the VL1 (Variable-valued Logic system 1) notation, a multiple-valued logic attribution calculus with typed variables. A *selector* is an attribute value condition that relates an attribute to a value or a disjunction of values. A *conjunction* of selectors is called a *complex* . A hypothesis is a disjunction of complexes that together describe all of the positive examples and none of the negative ones.

When building a decision rule, AQ15c performs a heuristic search through the space of logical expressions to determine those that account for all positive and no negative examples. Because there are usually many such strong rules, the goal of AQ15c is to find the most preferred one. This preference criterion can be defined by the user (the default criterion is chosen in this project). The weights associated with the generated rules (see the experiment result part) describe the coverage of individual rules. The first weight indicates the number of records which are satisfied by this rule. The second weight indicates the number of records which are satisfied by this rule only.

Another feature of AQ15c is its ability to handle inconsistent data and noisy data. AQ15c provides three options to treat inconsistent examples either as positive examples, negative examples or simply neglects them. (Inconsistent examples are treated as negative examples in this project). When the training set contains incorrect examples, it may be advantageous to apply the constructed rules in a probabilistic manner. If flexible matching is used, it is possible to simplify a description by removing one or more complexes, so called rule truncation. If the training set is noisy, the removed complexes may be indicative of errors in the data.

(6) Data mining

The results generated from step 5 are converted into the training set format and input into the AQ15c system. The system output is the classification rules to characterize El-Nino events and non El-Nino events. (see appendix for the training set and system output).

(7) Interpretation

These previously generated rules have been interpreted and some interesting patterns have been found. The results are shown and discussed in the later part of this paper. The interpretation rules are defined as follows:

a. Only the top three rules (from each class) are considered. The order of the rules is determined by the weights associated with them.

b. Only the "common" attributes (shown up in ALL event classes) are considered.

c. Only the outstanding "consistent" (cross ALL event classes) attributes are considered, for example, [VMAF=1..2] in El-Nino event class is consistent with [VMAF=4..5] in non El-Nino event class. The consistency should also be outstanding, e.g. [MAF=4] in El-Nino event class vs. [MAF=3] in non El-Nino event class can not be considered outstanding. Another observation is that every 'consistent' attribute is also a 'common' attribute.

d. If any conflicts occur, rules with higher weights have higher priority.

e. Based on the associated value ranges of considered attributes (from rule a, b, c, and d), candidate hypotheses are created for further studies.

3. Geophysical Datasets

Almost all the datasets used in this project are obtained from the NASA Goddard DAAC Climatology Interdisciplinary Data Collection. Each dataset specifies one of the four geophysical parameters (NDVI, temperature deviations, solor irradiance, China average summer rainfall index) for selected spatial units or areas of the globe.

3.1. Normalized Difference Vegetation Index (NDVI)

A number of studies have shown that Normalized Difference Vegetation Index (NDVI) provides an effective measure of photosynthetically active biomass (Asrar, et al, 1989). NDVI has also been shown to be well correlated with climate variables including rain fall and evapotranspiration in a wide range of environmental conditions. NDVI may therefore be considered to represent the integration of climate variables at a given location and time. Thus, NDVI is chosen as the major climate parameter to be considered in studying the teleconnections with El Nino phenomenon in this project.

The data set specifies values of NDVI for the whole Earth. It was produced as part of the NOAA/NASA Pathfinder AVHRR Land Advanced Very High Resolution Radiometers (AVHRR) on the "afternoon" NOAA operational meteorological satellites (NOAA-7, -9, -11), the data cover the period from July 1981 to August 1994.

On the NOAA-7, NOAA-9, and NOAA-11 satellites, the AVHRR sensor measures emitted and reflected radiation in five channels (bands) of the electromagnetic spectrum. The first AVHRR channel is in a part of the spectrum where chlorophyll causes considerable absorption of incoming radiation, and the second channel is in a spectral region where spongy mesophyll leaf structure leads to considerable reflectance. This contrast between responses of the two bands can be shown by a ratio transform. i.e. Normalized Difference Vegetation Index (NDVI), which has been shown to be highly correlated with vegetation parameters such as green-leaf biomass and green-leaf area and, hence, is of considerable value for vegetation coverage.

Characteristics of this dataset

The dataset contains values of NDVI for all land squares (approximately 100 km X 100 km) for each month in the period July 1981 to August 1994.

Parameters: Normalized Difference Vegetation Index (NDVI), derived from the visible and near-infrared channel reflectance (0.58 to 0.68 um and 0.73 to 1.10 um, respectively). The NDVI is highly correlated with surface vegetation.

Units: Dimensionless

Typical Range: -0.200 to 0.730

Temporal Coverage: July 1981 to August 1994

Temporal Resolution: All gridded values are monthly composites.

Spatial Coverage: Global

Spatial Resolution: 1 degree x 1 degree (approximately 100 km X 100 km)

3.2. Temperature Deviations

Monthly Surface air temperature anomalies for the period 1984-1994 have been calculated by the Climate Research Unit (CRU) of the University of East Anglia, Norwich, England, using data from several sources. The anomalies consist of land and ocean temperature departures (seasonal variations have been removed) and are given on a 5 x 5 degree world grid. The data from several sources were carefully examined and corrections were made to compensate for known measurement problems.

Characteristics of this dataset:

Parameters: Surface Temperature deviations

Units: Degrees Celsius

Typical Range: 2 degrees to 2 degrees Celsius

Temporal Coverage: 1984 - 1994

Temporal Resolution: Gridded monthly means

Spatial Coverage: Global

Spatial Resolution: 5 degree x 5 degree

3.3. Solar Irradiance

This dataset contains measurements of the monthly solar irradiance on Earth for the period of 1978 – 1993.

Total solar irradiances provides the energy source that powers the Earth's climate and biosphere. Data are derived from The Nimbus-7 Earth Radiation Budget (ERB) measurements (November 1978–December 1993).

Characteristics of this dataset:

Parameters: Total solar irradiance (for the whole earth)

Units: Watts/m^2

Typical Range: 1368.50 to 1374.80

Temporal Coverage: ERB November 16, 1978 through December 13, 1993

Temporal Resolution: monthly means

Spatial Coverage: This is satellite orbital data Spatial Resolution: N/A

3.4. China Average Summer Rainfall Index

The dataset provides a weighted average of China summer rainfall for the period of about 44 years. Data are normalized, Dimensionless. The temporal coverage is from 1951 to 1994, and the data is six to eight months average.

4. Experiment

The goal of this study is to derive hypotheses on El Nino teleconnections from the geophysical data described earlier. The output attribute is the presence or absence of El Nino. The selected input attributes for the training set used in this study are listed in Table 1.

Geophysical Attribute (Parameter)	*Alias*	*Latitude Range*	*Longitude Range*
NDVI of Northeastern coastal area of USA	VNEU	48 – 41	-78 – -65
NDVI of Middle Africa area	VMAF	11 – 1	7 – 43
NDVI of Southeastern coastal area of China	VSEC	37 – 21	112 -122
NDVI of Southwestern coastal area of Africa	VSWA	-3 – -23	32 – 42
T.D. of Northeastern coastal area of USA	NEU	48 – 41	-78 – -65
T.D. of Middle Africa area	MAF	11 – 1	7 – 43
T.D. of Southeastern coastal area of China	SEC	37 – 21	112 -122
T.D. of Southwestern coastal area of Africa	SWA	-3 – -23	32 – 42
T.D. of Southeastern coastal area of USA	SEU	43 – 26	-84 – -76
T.D. of Southern Pacific Region II	EL2	0 – -5	-90 – -80
T.D. of Central Pacific Region III	EL3	5 – -5	-150 – -90
China Summer Rainfall Index	CR	NA	NA
Solar Irradiance	SR	NA	NA

Table 1. Attributes used in the training examples

A brief description for each parameter can be found in section 3 (geophysical datasets). The number of discrete binned levels and data value ranges are listed in Table 2.

Geophysical Parameter	*N. of Levels*	*Data Value Range*	*Unit*
NDVI	5	-0.1333 – 0.4962	Dimensionless
Temperature Deviation (T.D.)	5	-6.825 – 9.895	Degrees C
China Average Rain Fall Index	3	-1.680 – 4.007	Dimensionless
Solar Irradiance	3	1370.8 – 1373.5	$Watts/m^2$

Table 2. Number of discrete values (levels) used for each attribute

Figure 1. Hypotheses generated by the AQ15c Learning Program

(Underlined attribute-value conditions were used for proposing hypotheses).

El Nino-outhypotheses:	
1 [EL3=2..3] [SWA=3..4] [$\underline{VNEU = 3..4}$][$\underline{VMAF = 1..2}$] [VSWA=2,4]	(t:6, u:5)
2 [SEC=3] [MAF=4] [VNEU=3..4] [$\underline{SR = 2..3}$]	(t:5, u:4)
3 [SEC=2] [SWA=3] [VSWA=1..3] [SR=3]	(t:4, u:3)
4 [EL2=2..3] [VNEU=4] [VSEC=1] [VMAF=5]	(t:3, u:3)
5 [NEU=3..4] [MAF=5] [VSWA=1..2]	(t:3, u:3)
6 [NEU=1] [SWA=3] [VSEC=1..2]	(t:3, u:3)
7 [SWA=3..4] [VNEU=5] [VMAF=4]	(t:3, u:2)
8 [EL3=2] [EL2=3] [VNEU=5] [VSWA=2..3]	(t:2, u:2)
9 [NEU=3,5] [VNEU=3] [VSEC=3..4] [VSWA=2..3] [CR=1]	(t:2, u:2)
10 [MAF=3] [CR=2] [SR=3]	(t:1, u:1)
11 [SEU=3] [NEU=2] [VNEU=4] [VSWA=4]	(t:1, u:1)

Non El Nino-outhypotheses:	
1 [SEC=1..4] [$\underline{VNEU = 1..2}$] [VSEC=3..5]	(t:18, u:14)
2 [MAF=3..5] [VSEC=2] [$\underline{SR = 1}$]	(t:15, u:13)
3 [EL3=2..3] [MAF=3] [VNEU=4] [$\underline{VMAF = 4..5}$]	(t:12, u:6)
4 [SEU=1..2] [VSEC=3..4] [VMAF=3..5]	(t:11, u:7)
5 [VNEU=1,3..4] [VMAF=5] [SR=2]	(t:8, u:3)
6 [MAF=2] [SWA=2] [VNEU=3..5] [VSEC=2,4]	(t:7, u:6)
7 [MAF=4..5] [VNEU=2..3] [SR=2]	(t:7, u:5)
8 [SWA=3] [VMAF=5] [VSWA=1]	(t:5, u:1)
9 [MAF=3..4] [VNEU=5] [VSEC=1] [VMAF=5]	(t:4, u:4)
10 [EL2=3..4] [VNEU=2..3] [VSWA=3..5] [SR=3]	(t:3, u:2)
11 [EL3=2] [EL2=4..5] [SEC=3] [VSEC=1]	(t:2, u:2)
12 [VNEU=2] [CR=1] [SR=3]	(t:2, u:1)

In our experiment, two events (El-Nino event vs. non El-Nino event) are used by the AQ15 system for exploratory hypothesis building.

The rules generated by the AQ15c are shown above. Based on the interpretation rules (see section 2), particularly interesting conditions are shown as underlined. The experiment input file (training set) and detailed experiment output (experiment results) can be found in the Appendix.

¿From the selected conditions (attributes and associated values, see Fig. 1 above), five hypotheses were subsequently derived:

(1) Vegetation coverage in an area centered in middle Africa is anti-correlated with El Nino signals ([VMAF=1..2] in El Nino-outhypotheses vs. [VMAF=4..5] in Non El Nino-outhypotheses).

The vegetation coverage in the Sahel zone and middle Africa has been an area of particular concern due to the associated repeated long-term droughts and changes in the land use patterns. There are a lot of climate related social-economical impacts, such as severe famines in those countries. Much work has been carried out to try to understand the relationship between changes in vegetation coverage and climate variations. There is evidences (Anyamba, et al, 1996) shown that El Nino signals shift the precipitation pattern, which is the most significant factor influencing the vegetation growth (normally one has an anti-correlation with the El Nino signals) due to semi-acid land type, through the teleconnection process. Because of the social-economical significance, more work is still on-going trying to understand the vegetation variations in those areas.

(2) Vegetation coverage in an area centered in the northeastern America coastal area may be correlated with El Nino signals ([VNEU=3..4] in El Nino-outhypotheses vs. [VNEU=1..2] in Non ElNino-outhypotheses).

One interesting fact is that the only significant El Nino impacted area in the North hemisphere is North America. There are a lot of weather related disasters reported relating with El Nino signals (Lau, et al. 1993). The effect of El Nino teleconnections in North America is currently the general belief, specifically the anti-correlated relationship between vegetation coverage variation and El Nino signals. However, one current on-going research suggests that there may exist some correlation relationships (positive relationship) in the northeastern coastal mixed forest area.

(3) There is a dependency between El Nino signals and variations of solar radiation ([SR=2..3] in El Nino-outhypotheses vs. [SR=1] in Non El Nino-outhypotheses).

This is an interesting hypothesis, because the time scales of the two signals are not matching each other: The solar radiation is quite stable, or in other words, variation time scales are longer (11 years). There is no obvious explanation for the relationship found here.

(4) There is no strong relationship between temperature deviations and El Nino signals.

This is understandable due to the fact that there are multiple forces influencing temperature variations, such as local circulation patterns, and the teleconnections with El Nino signals, even though significant in some regions, may be masked by other factors.

(5) There is no strong dependency between El Nino signals and China summer rainfalls

The major forces influencing China summer precipitation variations are Chinese and India monsoon events. The teleconnections with El Nino signals are believed to be a second order effect.

The above hypotheses represent interesting results of the work described here. Their validity, however, needs to be tested by collecting appropriate observational or modelling data.

5. Discussion

Among the basic problems facing Earth scientists are what kind of research hypotheses can be derived from the body of existing data and which hypotheses should be investigated in subsequent research to derive an acceptable theory. There are two traditional approaches to hypothèsis searching, namely modeling approach and statistical approach.

The basic assumption for the modeling approach is that the natural phenomenon can be explained by known natural laws. Based on this assumption, mathematical and physical models are developed first, initial and boundary conditions are set up, and then models are run. The advantage of this approach is that all inferences are valid assuming a validity of mathematical and physical laws. However, there are three major drawbacks in this approach. First, models may be incomplete. Nature is so complex that it is impossible for a given model to clearly describe all related natural processes. Second, the initial and boundary conditions can not be set up perfectly. Errors are unavoidable. Third, even if we have a perfect model and perfect initial and boundary conditions, because of the chaotic behavior of the underlying weather system, the system could be unstable and have no fixed final state. Thus, direct hypothesis inference based on physical and mathematical laws may not work in so! me situations, e.g. for El Nino teleconnections.

The other approach is to apply statistical exploratory data analysis to find promising hypotheses, such as the grand tours (Wegman, et al, 1992) in hyper-dimensional space. One problem for doing this is that no clear criteria to discriminate between candidate hypotheses. Another common problem for the statistical inference approach is that the underlying models need to be assumed first. However, normally no such model can be assumed in advance.

¿From the above discussions, we can see that neither direct inference through physical and mathematical laws nor statistical inference work very well in searching promising hypotheses. Thus, a machine learning & data mining approach is proposed here. Machine learning algorithms/systems, such as AQ15c, can be applied to data to generate hypotheses.

6. Conclusions

The advance of satellite and romote sensing technology provides Earth scientists a new way to understand the global climate system from space. However, one major problem facing Earth scientists is how to efficiently utilize the large volumes of satellite data to get a much more clear understanding of nature.

Data mining has been shown to be a promising approach to solve this problem. In this project, we apply the AQ15c system, an inductive machine learning system, to mine the satellite geophysical data and to generate interesting hypotheses. While no testing has been made based on the learned rules due to limited data, expert domain knowledge is instead applied to evaluate hypotheses. On the other hand, because of the extreme complexity of the natural system, attempts to get reasonably correct results based on traditional testing are almost impossible.

The data mining experiments reported here are at an early, exploratory stage. Due to limited time, only a small subset of geophysical parameters and associated data sets have been examined. More work is necessary, as described below:

(1) Researches should be carried out to test the generated hypotheses.

(2) A multistrategy approach to data mining may be considered in the future (Michalski and Kaufman, 1997) in order to extend the capabilities of the current approach.

(3) Since data mining is time consuming, more automatic and efficient ways need to be explored, especially in producing multi-dimensional and multi-scale data summary. A multi-dimensional data model needs to be developed.

(4) Database technology may be applied to efficiently index, retrieve, and process the large volume of scientific data.

References

Anyamba, A. & J.R. Eastman, "Inter-annual variability of NDVI over Africa and its relation to El Nino Southern Oscillation", INT. J. Remote Sensing, 1996, Vol.17, No.13, 2533-2548.

Asrar, G.(ed), "Theory and Applications of Optical Remote Sensing", John Wiley & Sons, 1989.

Deser, C. and J.M. Wallace, "El Nino Events and their Relation to the Southern Oscillation", J. Geophys, Res. 92(C13) 14189-14196.

Fayyad, U., D. Haassler & P. Stolorz, "Mining Scientific Data", Communication of the ACM, Vol.39, No.11, 1996.

Harries, J.E., "Earthwatch: The Climate from Space", Praxis, 1994.

Kafatos, M., Z. Li, R. Yang, X.S. Wang, et al. "The Virtual Domain Application Data Center: Serving Interdisciplinary Earth Scientists", Proceeding of Ninth International Meeting on Scientific and Statistical Database Management, IEEE, 1997.

Lau, K.M. & A.J. Busalacchi, "El Nino Southern Oscillation", Atlas of Satellite Observations related to Global Change, Cambridge University Press, 1993.

Michalski. R.S., I. Mozetic, J. Hong, N.Lavrac: "The AQ15 Inductive Learning System: An Overview and Experiments", ISG86-20, UIUCDCS-R-86-1260, Department of Computer Science, University of Illinois. Urbana, 1986.

Michalski, R.S., & Kaufman, K.K., "Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach, to be appear in *Machine Learning and Data Mining: Methods and Applications*, John Wiley & Sons Ltd,. 1997.

Vallis, G.K., J. Geophys. Res. 93(C11) 13979-13991, 1988.

Wegman, E.J. & D.B. Carr, "Statistical Graphics and Visualization", Center for Computational Statistics, George Mason University, 1992.

Wnek, J., Kaufman, K. Bloedorn, E. and Michalski, R.S., "Inductive Learning System AQ15c: The Method and User's Guild", Reports of the Macchine Learning and Inference Laboratory, MLI 95-4, George Mason University, Fairfax, VA22030, March, 1995.

Appendix A.1: Experiment Results

```
# See the report for the interpretation rules
# The interested attributes are listed in the following:
# 1. VMAF
# 2. VNEU
# 3. SR
# Interesting results are lightlighted in the generated rules.
# The values associated with the attributes stand for the corresponding
# discretized level (a higher level corresponding to a higher value)
#
#


   parameters
   run   mode   ambig   trim   wts   maxstar   echo    criteria  verbose
    1    ic     neg     mini   cpx    10       pdnv     default     1


   domaintypes
     type    size    cost    xthres    name
     lin       6     1.00              EL3
     lin       6     1.00              EL2
     lin       6     1.00              SEU
     lin       6     1.00              NEU
     lin       6     1.00              SEC
     lin       6     1.00              MAF
     lin       6     1.00              SWA
     lin       6     1.00              VNEU
     lin       6     1.00              VSEC
     lin       6     1.00              VMAF
     lin       6     1.00              VSWA
     lin       4     1.00              CR
     lin       4     1.00              SR


   variables
     #   type    size    cost    xthres    name
     1   lin      6      1.00              EL3.EL3
     2   lin      6      1.00              EL2.EL2
     3   lin      6      1.00              SEU.SEU
     4   lin      6      1.00              NEU.NEU
     5   lin      6      1.00              SEC.SEC
     6   lin      6      1.00              MAF.MAF
     7   lin      6      1.00              SWA.SWA
     8   lin      6      1.00              VNEU.VNEU
     9   lin      6      1.00              VSEC.VSEC
    10   lin      6      1.00              VMAF.VMAF
    11   lin      6      1.00              VSWA.VSWA
    12   lin      4      1.00              CR.CR
    13   lin      4      1.00              SR.SR


   El_Nino-outhypo
     #   cpx
     1   [EL3=2..3] [SWA=3..4] [VNEU=3..4] [VMAF=1..2] [VSWA=2,4]   (t:6, u:5)
     2   [SEC=3] [MAF=4] [VNEU=3..4] [SR=2..3]   (t:5, u:4)
     3   [SEC=2] [SWA=3] [VSWA=1..3] [SR=3]   (t:4, u:3)
     4   [EL2=2..3] [VNEU=4] [VSEC=1] [VMAF=5]   (t:3, u:3)
     5   [NEU=3..4] [MAF=5] [VSWA=1..2]   (t:3, u:3)
     6   [NEU=1] [SWA=3] [VSEC=1..2]   (t:3, u:3)
     7   [SWA=3..4] [VNEU=5] [VMAF=4]   (t:3, u:2)
     8   [EL3=2] [EL2=3] [VNEU=5] [VSWA=2..3]   (t:2, u:2)
     9   [NEU=3,5] [VNEU=3] [VSEC=3..4] [VSWA=2..3] [CR=1]   (t:2, u:2)
    10   [MAF=3] [CR=2] [SR=3]   (t:1, u:1)
    11   [SEU=3] [NEU=2] [VNEU=4] [VSWA=4]   (t:1, u:1)
```

```
NonEl_Nino-outhypo
  #   cpx
  1   [SEC=1..4] [VNEU=1..2] [VSEC=3..5]   (t:18, u:14)
  2   [MAF=3..5] [VSEC=2] [SR=1]   (t:15, u:13)
  3   [EL3=2..3] [MAF=3] [VNEU=4] [VMAF=4..5]   (t:12, u:6)
  4   [SEU=1..2] [VSEC=3..4] [VMAF=3..5]   (t:11, u:7)
  5   [VNEU=1,3..4] [VMAF=5] [SR=2]   (t:8, u:3)
  6   [MAF=2] [SWA=2] [VNEU=3..5] [VSEC=2,4]   (t:7, u:6)
  7   [MAF=4..5] [VNEU=2..3] [SR=2]   (t:7, u:5)
  8   [SWA=3] [VMAF=5] [VSWA=1]   (t:5, u:1)
  9   [MAF=3..4] [VNEU=5] [VSEC=1] [VMAF=5]   (t:4, u:4)
 10   [EL2=3..4] [VNEU=2..3] [VSWA=3..5] [SR=3]   (t:3, u:2)
 11   [EL3=2] [EL2=4..5] [SEC=3] [VSEC=1]   (t:2, u:2)
 12   [VNEU=2] [CR=1] [SR=3]   (t:2, u:1)


This learning used:
  System time:      1.540 seconds
  User time:        3.00  seconds
```

Appendix A.2: Training Set

```
parameters
  run   mode   ambig   trim   wts   maxstar   echo   criteria   verbose
   1    ic     neg     mini   cpx    10       pdnv   default      1


variables
    #   type     levels    name
    1   lin        6       EL3
    2   lin        6       EL2
    3   lin        6       SEU
    4   lin        6       NEU
    5   lin        6       SEC
    6   lin        6       MAF
    7   lin        6       SWA
    8   lin        6       VNEU
    9   lin        6       VSEC
    10  lin        6       VMAF
    11  lin        6       VSWA
    12  lin        4       CR
    13  lin        4       SR
```

El_Nino-events

EL3	EL2	SEU	NEU	SEC	MAF	SWA	VNEU	VSEC	VMAF	VSWA	CR	SR
2	4	4	4	3	4	4	3	3	2	4	1	1
2	2	2	2	2	3	3	3	4	2	4	1	1
2	3	3	2	3	3	3	4	3	3	4	1	1
2	2	2	2	2	2	3	5	2	4	3	1	1
2	2	2	2	2	2	2	4	1	5	2	1	1
2	3	3	3	3	4	4	5	1	4	1	1	1
2	4	4	4	4	5	5	4	1	5	1	1	1
2	3	3	3	3	4	4	4	1	5	1	1	1
2	3	3	3	2	3	3	3	3	5	2	1	1
2	1	1	1	3	2	3	2	2	4	3	1	1
2	3	3	3	2	3	3	3	4	2	4	1	2
3	3	3	3	3	3	3	4	5	2	4	1	3
3	3	3	3	3	4	4	4	3	3	5	1	3
2	3	3	3	3	4	4	4	3	4	3	1	2
2	3	3	3	3	3	3	5	2	5	3	1	3
2	3	3	3	3	3	3	5	2	5	2	1	2
3	3	3	3	4	5	4	5	1	5	1	1	3
1	1	1	1	3	3	3	4	2	5	2	1	3
1	2	2	2	3	3	3	4	1	5	3	1	3
2	2	2	2	2	3	3	3	1	5	3	1	3
1	2	2	2	3	3	3	2	2	4	3	2	3
2	3	3	3	5	5	5	2	4	2	2	2	3
3	3	3	3	3	4	4	3	5	1	2	1	3
5	5	5	5	2	5	5	3	4	1	3	1	3
2	2	2	2	2	3	3	3	3	2	2	1	2
2	2	3	2	2	3	3	4	2	3	2	1	3
2	3	3	3	3	4	4	4	2	4	2	1	2
2	3	3	3	2	3	3	5	1	4	1	1	3
1	2	2	2	2	3	3	4	2	4	1	1	3
1	1	1	1	3	3	3	4	1	4	1	1	3
3	4	4	4	3	4	4	3	2	4	2	1	3

NonEl_Nino-events

EL3	EL2	SEU	NEU	SEC	MAF	SWA	VNEU	VSEC	VMAF	VSWA	CR	SR
2	3	3	3	3	4	4	4	2	4	1	1	1
2	3	3	3	2	3	3	4	1	4	1	1	1

2	3	3	3	3	4	4	4	2	5	1	1	1
2	1	1	1	3	3	3	4	3	5	1	1	1
2	3	3	3	4	4	4	2	3	5	2	1	1
2	1	1	1	1	1	1	1	4	3	2	2	1
2	5	5	5	4	5	5	1	5	1	3	2	1
2	3	3	3	3	4	3	1	4	1	3	1	2
2	3	3	3	3	4	4	2	5	1	4	1	1
2	2	2	2	3	3	3	3	4	3	3	1	1
2	3	3	3	3	4	4	4	2	4	2	1	1
2	3	3	3	3	4	4	5	2	5	1	1	1
2	1	1	1	2	2	2	5	3	5	1	1	1
2	2	2	2	2	2	2	5	2	4	1	1	1
2	3	3	3	3	4	4	4	2	5	1	1	1
2	3	3	3	3	4	4	3	2	5	2	1	1
2	2	2	2	3	3	3	2	2	5	4	1	1
2	4	4	4	3	5	4	2	2	4	3	2	1
2	4	4	4	1	3	3	2	2	2	3	2	1
2	1	1	1	2	2	2	1	3	4	3	2	1
2	3	3	3	3	4	4	1	3	3	3	2	1
2	4	4	4	2	4	4	2	4	1	4	1	1
2	1	1	1	2	2	2	3	4	2	4	1	1
2	1	1	1	2	2	3	4	3	3	4	1	1
2	2	2	2	2	3	3	4	2	4	3	1	1
2	2	2	2	2	3	3	4	1	4	2	1	1
2	3	3	3	3	4	4	5	1	5	1	1	1
2	2	2	2	2	3	3	4	2	5	1	1	1
2	4	4	4	3	4	4	4	1	5	1	1	1
2	2	2	2	3	3	3	3	2	5	3	2	1
2	2	2	2	3	3	3	2	2	4	3	2	1
2	2	2	2	4	4	4	1	2	3	4	3	1
2	1	1	1	2	4	3	1	3	2	3	3	1
2	1	1	1	3	1	1	1	4	1	4	2	1
2	1	1	2	2	3	3	2	4	2	4	2	2
2	2	2	2	3	3	3	3	3	3	3	1	2
2	2	2	2	1	2	2	4	2	4	3	1	2
2	2	2	2	2	3	3	5	1	5	2	1	2
2	3	3	3	3	4	4	5	1	5	1	1	2
2	2	2	2	2	2	2	5	2	5	1	1	2
2	5	5	5	3	5	5	4	1	5	1	1	1
2	3	3	3	2	4	3	3	1	5	1	2	2
2	2	2	2	3	3	3	1	1	5	2	2	2
2	3	3	3	3	4	4	2	1	3	2	3	2
2	3	3	3	3	3	3	2	2	1	2	3	1
2	4	4	4	3	5	4	1	4	1	4	2	2
2	3	3	3	3	4	4	1	3	1	4	2	2
2	4	4	4	3	5	5	3	2	2	4	1	2
2	3	3	3	2	3	3	4	2	4	4	1	2
2	2	2	2	2	3	3	4	2	5	3	1	2
2	2	2	2	1	2	2	5	2	4	2	1	2
3	3	3	3	2	3	3	4	2	4	2	1	2
2	3	3	3	3	3	3	4	2	5	1	1	2
1	2	2	2	2	2	2	3	1	5	3	1	2
2	2	2	2	1	2	2	4	2	5	4	1	2
2	3	3	3	1	2	2	2	4	4	4	2	1
4	4	4	4	3	4	4	2	4	3	4	2	2
4	4	4	4	4	5	5	3	5	2	4	1	2
3	3	3	3	4	4	4	3	5	3	5	1	2
1	2	2	2	3	3	3	4	3	3	5	1	3
2	3	3	3	3	3	3	4	3	4	4	1	3
3	3	3	3	4	4	4	4	2	5	3	1	2
3	3	3	3	3	4	4	5	1	5	3	1	3
2	3	3	3	3	3	3	5	2	5	1	1	3
2	3	3	3	2	3	3	4	2	5	3	1	2
2	2	2	2	3	3	3	4	2	5	3	1	3
1	1	1	1	2	2	2	3	2	5	3	1	3
4	4	4	4	2	4	4	2	3	4	2	2	3

4	4	4	4	3	5	4	3	4	3	3	2	3
1	2	2	2	3	3	3	2	2	4	3	1	3
1	2	2	2	3	4	4	1	3	3	2	2	3
2	2	2	2	3	4	4	1	5	2	3	2	3
2	3	3	3	5	3	4	2	5	1	4	1	3
3	3	3	3	2	3	3	3	5	2	5	1	3
4	4	4	4	2	5	4	3	5	2	5	1	2
1	2	2	2	3	3	3	4	4	4	4	1	2
2	2	2	2	2	2	2	4	3	5	3	1	3