An Event Detection Framework for Virtual Observation System: Anomaly Identification for An Acme Land Simulation

Zhuo Yao[1], Dali Wang[2] *, Yifan Wang[1], and Fengming Yuan[2]

[1] Department of Electric Engineering and Computer Science, University of Tennessee, TN 37996 USA,
[2] Environmental Science Department, Oak Ridge National Laboratory, TN 37831, USA,
* Corresponding author: wangd@ornl.gov

Abstract. Based on previous work on in-situ data transfer infrastructure and compiler-based software analysis, we have designed a virtual observation system for real time computer simulations. This paper presents an event detection framework for a virtual observation system. By using signal processing and detection approaches to the memory-based data streams, this framework can be reconfigured to capture high-frequency events and low-frequency events. These approaches used in the framework can dramatically reduce the data transfer needed for in-situ data analysis (between distributed computing nodes or between the CPU/GPU nodes). In the paper, we also use a terrestrial ecosystem system simulation within the Earth System Model to demonstrate the practical values of this effort.

1 Introduction

Considerable effort has been made to develop accurate and efficient climate and Earth system simulations in the last two decades. Climate change analysis with both domain knowledge and observational datasets has drawn more and more attention since it seeks to assess whether extreme climate events are consistent with internal climate variability only, or are consistent with the expected response to different combinations of external forces and internal variability[10][12]. However, detecting extreme events in large datasets is a major challenge in climate science research. Current algorithms for detecting extreme events are founded upon scientific experience in defining events based on subjective thresholds of relevant physical variables[7]. dos Santos et al. proposes an approach to detect phenological changes through compact images[11]. Spampinato et al. propose an automatic event detection system based on the Makov Model[3]. Nissen et al. propose a technique for the identification of heavy precipitation events, but only by means of threshold identifications, which is not suitable for big database[7]. Gao et al. detect the occurrence of heavy precipitation events by using composites to identify distinct large-scale atmospheric conditions[9]. Zscheischler et al.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

present a methodological framework, also using thresholds, to detect spatiotemporally contiguous extremes and the likely pathways of climate anomalies[17]. Shirvani et al. develop and investigate a temperature detection model to detect climate change, but it is limited to a single domain[14]. The common theme in all of the above event detection methods is that it only considers post simulation data analysis. When analyses are performed in post-simulation mode, some or all of the data is transferred to different processors, either on the same machine or all together on different computing resources all together[4]. However, in reality, the data streams in climate simulations are enormous, which makes the data transfer over network unaffordable. In addition, with such enormous data streams, the memory and the calculating power of the remote machine would be rapidly exceeded. Furthermore, researchers can take action immediately based on the detected events while the system simulation is running and benefit most from the performance of graphics processing unit (GPU). We propose an unsupervised event detection approach that does not require human-labelled data as was required by [3][1]. This is an advantage since it is not clear how many labels are needed to understand events in a huge database. Instead of human labeling, we expect the infrastructure to learn bench patterns through long-term experiment datasets under an unknown background. For all these reasons, we propose an event detection framework for the virtual observation system (VOS) that provides run-time observation capability and in-situ data analysis. Our detection method enables our processing framework to detect events efficiently since the complexity of the output space is reduced. In this paper, we begin by introducing the VOS framework and then describe the functionalities of its components. Secondly, we explain how to apply signal-processing theory to reduce data and capture high and low frequency anomalies. Finally, we use the framework to identify anomalies and events then verify the detected events using observed datasets in Accelerated Climate Modeling for Energy (ACME) simulation.

2 Event Detection for Virtual Observation System

2.1 Virtual Observation System and Design Considerations

Over the past few decades, climate scientists and researchers have made tremendous progress in designing and building a robust hierarchy framework to simulate the fully coupled Earth system. This simulation can advance our understanding of climate evolution and climate extreme events at multiple scales. Significant examples of event information about extreme climate phenomena include floods[8], precise water availability, storms probability, sea level, the frequency and duration of drought, and the intensity and duration of the extreme heat. Understanding the role of climate extremes is of major interest for global change assessments; in addition, such phenomena have enduring and extensive influence on national economies. In detecting events in such a large dataset within the extreme-scale computing context, I/O constraints can be a great challenge. Scientists typically tolerate only minimal impact on simulation performance, which places significant restrictions on the analysis. In-situ analysis typically shares

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

primary computing resources with simulation and thereby encounters fewer resource limitations because the entirety of the simulation data is locally available. Therefore, a potential solution is to change the data analysis pipeline from post-process centric to a concurrent approach based on in-situ processing. Moreover, a GPU has a massively parallel architecture consisting of thousands of smaller, more efficient cores designed for handling multiple tasks simultaneously which accelerate analytics. The simulation only analyze variables status in real time. In stead, scientists and researchers want to know what elements increase/decrease abnormal immediately, therefore they would decide what action to take when what type of event hanppens. A previous paper[15] presented a virtually observed system (VOS) that provides interactive observation and run-time analysis capability through high-performance data transport and in-situ data process method during system simulation.

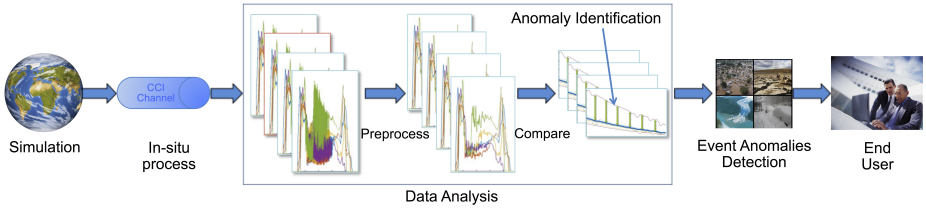

Fig. 1. VOS Overview.

Figure 1. illustrates how the VOS works. The VOS framework has three components: the first one is a compiler-based parser, which analyses target modules'internal data structure and inserts the data stream's statement to the original model code. The second component is the communication service using CCI (common communication interface), an API that is portable, efficient, and robust to meet the needs of network-intensive applications[2]. Once the instrumented scientific code starts to simulate, the VOS turns on the CCI channel to listen and interact with the simulation. The CCI channel employs a Remote Memory Access method to send remote buffers to the data analysis component in GPU through network since the parallelism of CPU is much lower than GPU[5]. The last component is data analysis, which collects and analyses data signals and then visualizes events for end-users. The first two components are explained in our previous work[15][6]. This paper will focus on presenting the event detection in data analysis component.

2.2 Data Reduction via Signal Processing

Within the VOS for climate simulation, the analysis component can potentially receive hundreds of variables every simulation timestep (half an hour) from every single function module. To deal with the I/O challenge presented by the enormous, periodic data transfer features, signal processing is proposed. Signal

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

processing is an enabling technology that encompasses the fundamental theory, applications, algorithms and implementations of processing or transferring information contained in many different physical, symbolic or abstract formats broadly designated as signals[6]. Because the memory and computation capability of the second resource is limited, the use of a lower sampling rate results in a implementation with less resource requirement. Nonetheless, downsampling alone causes signal components to be misinterpreted by subsequent users of the data. Therefore, for different science research requirements, different signal filter methods are needed to smooth the signal to an acceptable level. If researchers are interested in long period events result from multi physical elements anomalies, a low-pass filter can be used to remove the short-term fluctuations, and leave the longer-term trend through, since the low-pass filter only permits low-frequency signals and weakens signals with frequencies higher than the cutoff frequency. In contrast, if researchers are interested in abrupt change in a short time period, a filter can be used to pass high-frequency signals and weaken lower than cutoff frequency signals. Our data reduction process consists of two steps: first, a digital filter is used to pass low/high-frequency signal samplings and reduce high/low-frequency variable samplings and then the filtered signal sampling rate is decimated by an integer factor α, which means only keep every α th sample. Based on Nyquist sampling theorem, the sufficient α could be doubled or larger than the original frequency. Nyquist sampling theorem establishes a sufficient condition for a sample rate that permits a discrete sequence of samples to capture all the information from a continuous-time signal of finite bandwidth[13].

3 A Case Demonstration for Acme Land Model

This section reports a detailed event detection implementation and result verification for the ACME case. The ACME is a fully-coupled, global climate model that provides state-of-the-art computer simulations of the Earthrq's past, present, and future climate states. Within ACME, the ACME Land Model (ALM) is the active component to simulate surface energy, river routing, carbon cycle, nitrogen fluxes and vegetation dynamics[16].

3.1 ACME Land Model for NGEE Arctic Simulation

In this case study, ALM was configured as a single-landscape grid cell simulation conducted offline over Barrow, Alaska, the Next Generation Ecosystem Experiments Arctic site. The purpose of the case study was to investigate terrestrial ecosystem responses to specific atmospheric forcing. The ALM has three hierarchical scales within a model grid cell: the land unit, the snow/soil column, and the plant functional type (PFTs). Each grid cell has a different quantity of land units with various columns, and each column has multiple PFTs. For demonstration purposes, the observation system only tracks the variable flow of a CNAllocation module which has been developed to allocate key chemical elements of a plant (such as carbon, nitrogen and phosphorus) within a terrestrial ecosystem.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

3.2 Detection Framework

For the single CNAllocation module, the data flow includes three hundred variables. The NGEE simulation generates and sends out variables every half hour. The whole simulation period is 30 years, which means the data analysis component receives hundreds of multi-dimensional variables for 30*365*48=525600 times. To manage the huge quantities of data generated by the simulation, each of which had a large frequency, we employed frequency domain signal processing. The framework is schematically illustrated in Figure 2., which identifies anomalies of various durations and spatial extents in the Barrow Ecosystem Observatory (BEO) land unit datasets. In the first step, the framework filters out the interesting elements from the dimensional arrays and then apples decimation process to reduce the 30 years worth of variables. To find the average monthly pattern, only the first 6 years worth of data are initially selected. Once the monthly pattern for each variable is calculated from the training set, the framework proposes a detection algorithm based on Euclidean distance and compares the Euclidean distance the 30 years'data with the monthly pattern. If the normalized distance exceeds a threshold, the framework marks this variable in this month and this year as an anomaly alert. Finally, if the number of accumulated alerts in one year is very large, this time period is considered as an interesting event. Each detected event can consist of several patch boxes and can last for several time steps. Below is the detailed detection process.

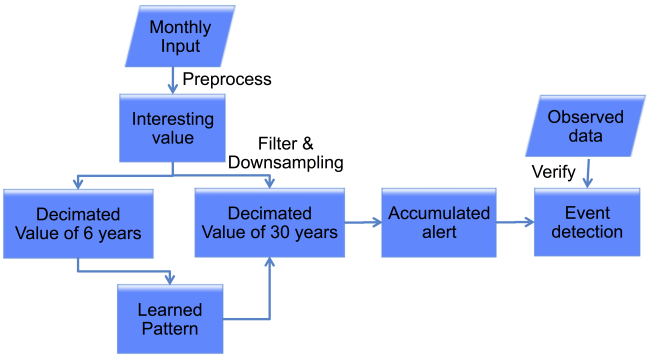

Fig. 2. Detection Framework. It first decimates 30 years' variables values, then uses first 6 years' data to find averaged monthly patterns, last tracks the Euclidean distance to find anomalies.

3.3 Event Detection

Variable Preprocess The climate change system defines, generates and calculates nutrient dynamics as the way they are in an ecosystem (build up, retain, transfer etc). In our work, the module CNAllocation has 320 nutrient dynamics

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

related variables, some of which are one-dimensional array, and some of which are two-dimensional array. For example, in $cnstate\_vars\%activeroot\_prof\_col$ (number of active root distributed through column), the first dimension denotes the column number and the second dimension stores the active root numbers for that relevant column. The variable $carbonstate\_vars\%leafc\_storage\_patch$ is a one-dimensional array with 32 elements that stand for the C storage in a leaf for every PFT level. The purpose of this step is to select out four elements from the default, since the BEO site only has four different plant types. Table 1. shows the indexes of these plant types and their meanings.

Table 1. Variable's PFT index meaning.

PFT Index	Meaning
0	Not vegetated plants
9	Shrub with broadleaf and evergreen
11	Boreal shrub with broadleaf and deciduous
12	Arctic grass with c3

Data Process To simultaneously save memory and retain as many of the data's contours as possible, the framework uses low-pass filter and down sampling data processing method. For example, the variable $carbonflux\_vars\%cpool\_to\_xsmrpool\_patch$ in year 1997, maintenance of respiration storage pool, the original values shown in the upper left panel of Figure 3. include all year-round (17520 timestep) value of a single variable. The size of these data requires around 0.07MB in disk space. The total store memory would be 672 MB if we catch and store all variables'information that is not necessary and burdensome for in-situ analysis. However, if the framework applies the data reduction method directly to the original dataset, the signal becomes aliased of original continuous signal, just as the information shown in the lower left panel of Figure 3. The first and third quarters information are phased out. In other words, whether the decimated signal information maintains the original features massively depends on which decimator the algorithm chooses. If the decimator reflects the variable's frequency, the output signal line will be similar to the original; otherwise, the signal line will change considerably. The framework applies low-pass filter first in consideration of long run trends and anomalies. The right two panels in Figure 3. represent the result of the low-pass method and the subsequent downsampling output, respectively, which together maintain the original features. In the experiment, the downsampling decimator $1/\alpha$ was set to 1/48, which eventually downsized the one-year variable's memory to 1.49KB for single timestep.

Pattern Estimation The framework estimates the monthly averaged pattern for every variable in each month (Jan-Dec) using the simulation data of the past six years'and gets 12*320=3840 bench month patterns in total. Every thin line in

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

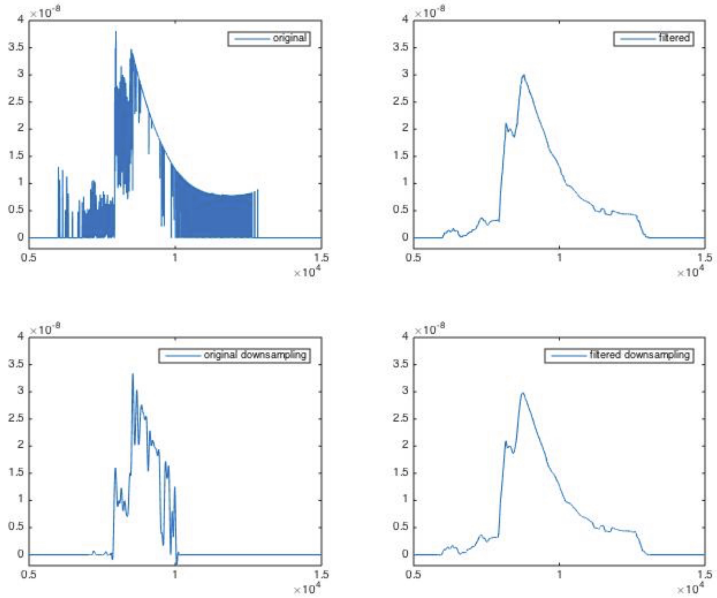

Fig. 3. Downsampling and interpolation. The left panel shows the result directly come from downsampled signals. The right panel shows result signals through filtering and downsampling, which is more accurate than left.

Figure 4. shows the value and pattern of July a conopyflux variable. The name of this variable is $CNCarbonFlux\%cpool\_to\_xsmrpool\_patch$, which represents the flux from total carbon pool to the maintenance respiration storage pool, and the thick blue line represents the averaged pattern of this variable in July.

Anomaly Identification Based on the monthly averaged patterns, we can compare the Euclidean distance between the data in each individual month and the monthly averaged pattern using:

$$D_i = \sqrt{\sum_t [X_i(t) - \bar{X}(t)]^2}, \tag{1}$$

$$\bar{X}(t) = avg[X_j(t)], j \in [i - N, i - 1] \tag{2}$$

The distance is normalized to get a more robust relationships to adjust values measurement from different scales to same scales and reduce the effect of data anomalies. Below is used to normalize every Euclidean distance to range in [0,1]:

$$\widetilde{D}_i = \left[\frac{D_i - \min_j D_j}{\max_j D_j - \min_j D_j} \right]^+, \tag{3}$$

$$j \in [i - N, i - 1] \tag{4}$$

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

Below is used to evaluate whether the variable of individual month becomes anomaly:

$$Alert = \begin{cases} 0 & \widetilde{\underset{i}{D}} > \gamma, \\ 1 & \widetilde{\underset{i}{D}} \leq \gamma. \end{cases} \tag{5}$$

If the normalized distance is larger than the set up threshold of value 0.8, the framework will flag the input simulation data streams as an interesting anomaly alert. Figure 4. shows the variable *cpool_to_xsmrpool_patch* of July 1992 is an extreme anomaly because the normalized distance is big.

Event Detection The framework identifies the entire anomaly for every single variable in every month of 30 years and records the total number of alerts in each month. Figure 5. displays accumulated alert count in 30 years with 320 variables. The overall anomaly peaks can be found in the monthly comparison curve and are accumulated among the year dots. Four extreme events were detected from the horizontal comparison. These events happened in May 1991, which had more than 120 alerts, October 2000, which has 180 alerts, Jun and Jul 1997 and Sep 1998 which had more than 100 alerts. From the vertical comparison, the year of 1997, 1998 and 2000 have the most alerts caused by extreme events. Based on this analysis, we can see that extreme weather events may take place in year 1997, 1998 from Jun to Sep and year 2000 from Jun to Nov. Further verification is needed to for the detection results. Furthermore, we need to investigate what kind of event occurred and the cause of those events.

3.4 Event Verification

In the last step, we verify the event through the input data and identify the event type. The climate experience tells us that temperature and precipitation are the top two factors that affect the results. Therefore, the two variables from year 1990 to 2000 were collected and analyzed. Figure 6. show the temperature at the beginning of December in year 1995 was high and the month had large temperature fluctuation. In year 1996, the temperature trend was similar to that of year 1995, but temperature was higher than any other years. These two curves explain the year 1996 had a warm winter that was part of an arctic warming trend. This trend is most observable during winter. Although most ecosystem activity is in dormancy in cold winter, soil microbial activity can still be significant especially if lasting or significant warming occurs. This includes enhanced soil heterotrophic respiration, methane generation, and nitrogen mineralization and its cascading reactions like nitrification and denitrification. The consequent Inorganic N accumulation during winter period can also cause large denitrification in early spring due to snow melting, which cause saturated soil conditions. Therefore, in the years 1997 and 1998, there was a great deal of variation among different variables, which caused many alerts. Figure 7. compares precipitation from year 1995 to 2000, showing that the daily precipitation

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

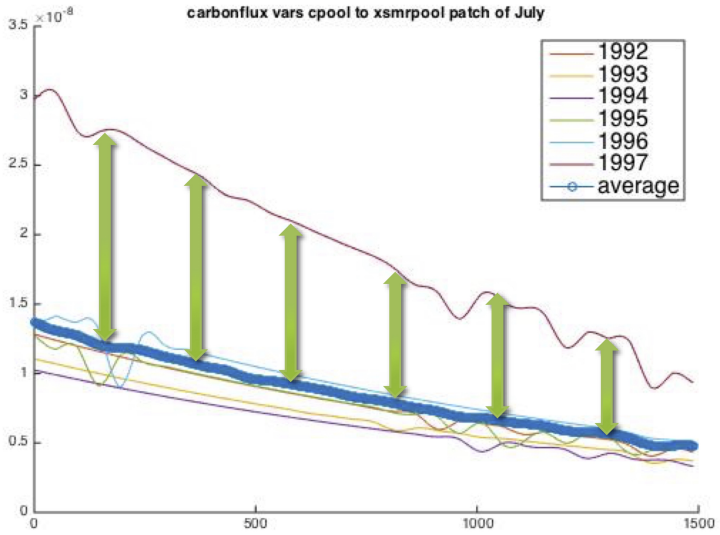

Fig. 4. July pattern comparison of variable *cpool_to_xsmrpool_patch* from year 1992 to year 1997. Among them, bold line is the July averaged pattern.

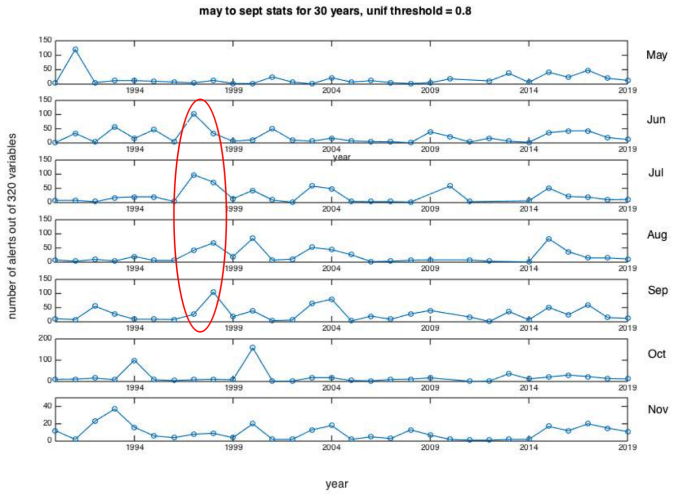

Fig. 5. Accumulated 320 variables anomaly alert count comparison from May. to Nov. in 30 years. Year 1997 and year 1998 have continuous events since the alert counts keep peak among all these years.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

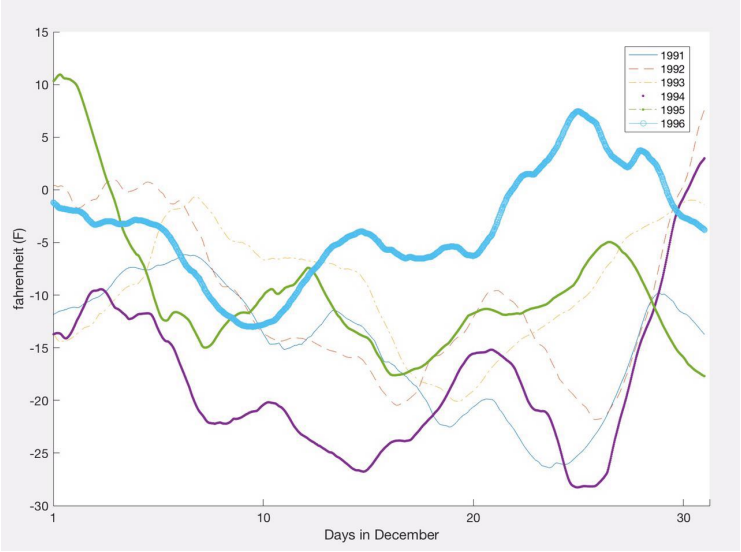

Fig. 6. December daily temperature in F from year 1991 to year 1996, which explains why year 1997 and year 1998 have more than 100 anomaly alerts. December daily temperature in the year 1996 was higher than any other years' and the warmer winter feature could also be reflected from Figure 5's November alert count. The warming trend therefore caused a great deal of variation among different variables in year 1997 and year 1998.

in Year 2000 was greater than that in the other years. Heavy precipitation or rainfall usually causes soil saturation (i.e. anaerobic conditions), which favors methane production, and N gaseous emission from mineralization, nitrification and especially denitrification. Extreme rainfall has a huge impact on spontaneous and large fluxes of greenhouse N gas and methane from soils. Therefore, the numbers of alerts are significant from July to November in Year 2000.

4 Conclusions

Climate change analysis of large datasets is time-consuming; in addition, the post-simulation processes that transfer tremendous data to other resources rapidly exceed the latter's memory and calculation power. In previous work, the virtual observable system with data flow analysis parser and in-situ communication infrastructure was proposed in previous work to analyze climate model data in real time. This paper presents an event detection analysis framework under the VOS. By using the decimation method in digital signal processing, the framework can reduce data transfer considerably and maintain most features of the original data. Through the event detection approach and the in-situ infrastructure, the framework can capture high frequency and low frequency anomalies, long-term extremes and abrupt events. It can also dramatically reduce pressure on remote

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

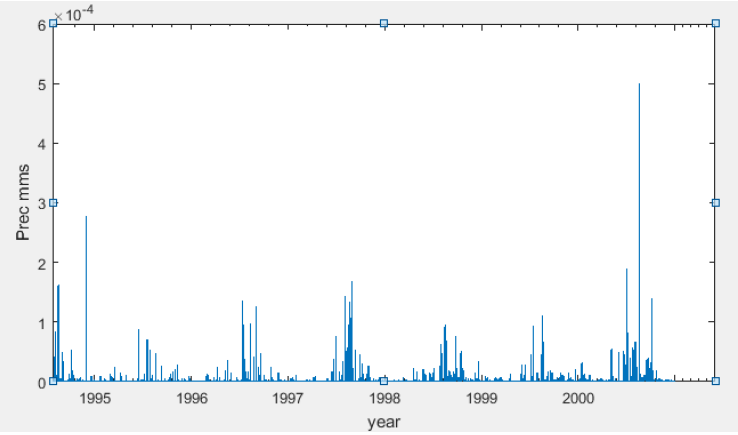

Fig. 7. Daily precipitation from years 1995 to year 2000.The precipitation in the second half of 2000 is heavier than any other years, which verify our detection result that from Jun to Nov, the total alert count is high due to the extreme rainfall's impact on spontaneous and large fluxes of greenhouse N gas and methane from soils.

processors. The practical values of this framework have been verified and demonstrated through the case study of a land model system simulation at BEO in Barrow, Alaska. In the future, after learned from the found patterns" features, we can use the variables collected from censors in the experiment combined with machine learning algorithms to predict the big event in advance.

ACKNOWLEDGEMENTS

This research was funded by the U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research (BER) program, and Advanced Scientific Computing Research (ASCR) program, and LDRD #8389. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

1. Aljawarneh, S., Aldwairi, M., Yassein, M.B.: Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science (mar 2017), http://linkinghub.elsevier.com/retrieve/pii/S1877750316305099
2. Atchley, S., Dillow, D., Shipman, G., Geoffray, P., Squyresz, J.M., Bosilcax, G., Minnich, R.: The Common Communication Interface (CCI). Proceedings - Symposium on the High Performance Interconnects, Hot Interconnects (Cci), 51–60 (2011)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4

3. Beauxis-aussalet, C.S.E., Palazzo, S., Beyan, C., Ossenbruggen, J.V., He, J., Boom, B., Huang, X.: A rule-based event detection system for real-life underwater domain pp. 99–117 (2014)
4. Bennett, J.C., Abbasi, H., Bremer, P.t., Grout, R., Gyulassy, A., Jin, T., Klasky, S., Kolla, H., Parashar, M., Pascucci, V., Pebay, P., Thompson, D., Yu, H., Zhang, F., Chen, J.: Combining In-situ and In-transit Processing to Enable Extreme-Scale Scientific Analysis
5. Du, P., Luszczek, P., Tomov, S., Dongarra, J.: Soft error resilient QR factorization for hybrid system with GPGPU. Journal of Computational Science 4(6), 457–464 (nov 2013), http://linkinghub.elsevier.com/retrieve/pii/S1877750313000161
6. Moura, J.: What is signal processing? [President's Message]. IEEE Signal Processing Magazine 26(6), 2009 (2009)
7. Nissen, K.M., Ulbrich, U.: Will climate change increase the risk of infrastructure failures in Europe due to heavy precipitation ? (October), 1–22 (2016)
8. Pitman, E.B., Patra, A.K., Kumar, D., Nishimura, K., Komori, J.: Two phase simulations of glacier lake outburst flows. Journal of Computational Science 4(1-2), 71–79 (jan 2013), http://linkinghub.elsevier.com/retrieve/pii/S1877750312000440
9. Program, C.J., Change, G., Noaa, I.E., Program, O.J., Change, G., Engineering, E.: An Analogue Approach to Identify Heavy Precipitation Events : Evaluation and Application to CMIP5 Climate Models in the United States pp. 5941–5963 (2014)
10. Santer, B.D., Mears, C., Doutriaux, C., Caldwell, P., Gleckler, P.J., Wigley, T.M.L., Solomon, S., Gillett, N.P., Ivanova, D., Karl, T.R., Lanzante, J.R., Meehl, G.A., Stott, P.A., Taylor, K.E., Thorne, P.W., Wehner, M.F., Wentz, F.J.: Separating signal and noise in atmospheric temperature changes : The importance of timescale 116, 1–19 (2011)
11. Santos, L.C.B., Almeida, J., Santos, J.A., Guimar, S.J.F., Ara, A.D.A., Alberton, B., Morellato, L.P.C., Torres, R.S.: Phenological event detection by visual rhythm dissimilarity analysis (2014)
12. Sciences, O., Sciences, E., Carolina, N., Group, C.D., Physics, P., Kingdom, U., Arbor, A.: Detection of Human Influence on a New , Validated 1500-Year pp. 650–667 (2006)
13. Shannon, C.: Editorial note on "Communication in the presence of noise". Proceedings of the IEEE 72(12), 1713–1713 (1984)
14. Shirvani, A., Nazemosadat, S.M.J., Kahya, E.: Analyses of the Persian Gulf sea surface temperature : prediction and detection of climate change signals pp. 2121–2130 (2015)
15. Wang, D., Yuan, F., Ridge, O., Pei, Y., Yao, C., Hernandez, B., Steed, C.: Virtual Observation System for Earth System Model : An Application to ACME Land Model Simulations 8(2), 171–175 (2017)
16. Yao, Z., Jia, Y., Wang, D., Steed, C., Atchley, S.: In Situ Data Infrastructure for Scientific Unit Testing Platform 1. Procedia Computer Science 80, 587–598 (2016), http://linkinghub.elsevier.com/retrieve/pii/S1877050916307591
17. Zscheischler, J., Mahecha, M.D., Harmeling, S., Reichstein, M.: Ecological Informatics Detection and attribution of large spatiotemporal extreme events in Earth observation data. Ecological Informatics 15, 66–73 (2013), http://dx.doi.org/10.1016/j.ecoinf.2013.03.004

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_4