Analysis of Climate Data in HPC Environment using Data Mining Approaches

Sangeetha M
Computer Science Department,
Vivekananda Institute of Technology,
Gudimavu, Kengeri Hobli,
Bangalore-74, India

N G Nivedita
Computer Science Department,
Vivekananda Institute of Technology,
Gudimavu, Kengeri Hobli,
Bangalore-74, India

Dr. K. C Gouda
CSIR Fourth Paradigm Institute
Wind Tunnel road, Bangalore-37,
India

Lakshmikantha G C
Computer Science Department,
Vivekananda Institute of Technology,
Gudimavu, Kengeri Hobli,
Bangalore-74, India

Abstract:- **On the growing importance of climate change studies and High Performance Computing, different users like farmer, scientist and policy maker needs to understand the various changes in the weather and climate parameters i.e., temperature, rainfall, humidity etc. Data discovery from temporal, spatial and spatiotemporal data is critical for climate science and to study the climate impacts on various sectors like health, water energy etc. Climate statistics is one of the mature area. The recent growth in observations and model outputs, combined with the increased availability of geographical data going to present new opportunities for data miners. In the present work an approach will be carried out to provide better understanding of the weather and climate data using spatiotemporal data mining. The challenges result from long-range to long-memory and possibly an nonlinear dependence, nonlinear dynamical behavior, presence of thresholds, importance of an extreme events or an extreme regional stresses caused by global climate change, uncertainty quantification, and the interaction of climate change with the natural and built environments. In this paper the different approaches (like data mining, numerical modeling etc.) of climate data analysis are explored and presented.**

Keywords: Climate Change, High Performance Computing, Data discovery, Spatiotemporal Data

1. INTRODUCTION

1.1 Climate change

Climate change is a change in the statistical distribution of weather patterns when that change lasts for an extended period of time. Climate change is a change in average weather conditions, or in the time variation of weather around longer-term average conditions. The climate of the planet has changed tremendously over the last few decades due to pollution, greenhouse gases and depletion of the ozone layer which protects the earth. Global warming is the major factors of climate change that leading to excessive flooding, forest fires and rise in global temperatures. Climate change been analyzed by scientists and some main features that work like early warning signs such as heat waves, periods of warm weather that are unusual, sea level rising, coastal flooding, ocean warming, and melting of glaciers are identified. In some of the areas, excess and unprecedented rise in temperature causes heat-related illness and death, especially in urban areas and majorly among the elderly, the young, the ill, and the poor. Different studies show it is very likely that hot extremes, heat waves, and heavy precipitation events will continue to become more frequent [1-2]. In addition to probable increase in intensity-duration-frequency (IDF) of extreme events and consequent exacerbation of natural hazards, it is also mentions that regional climate change is expected to cause stresses to the environment and society owing to increased temperatures and regional changes in precipitation patterns [1].

Now a day's climate changing and its need to study climate change because of the following reasons:

1. Studying climate change helps to understand what causes the changes.
2. Prepares us for any natural hazard or extreme changes that can be predicted.
3. Helps identify both man-made and natural causes for climate change.
4. Helps to understand how climate change has an impact on human health and the environment.

Climate data is very huge. So we need to implement climate data and climate model in the High Performance Computing environment.

2. HIGH PERFORMANCE COMPUTING (HPC)

High Performance Computing (HPC) is the generic name for the most powerful system available at the frontline of current processing capacity i.e., particularly speed of calculation. The term "SUPERCOMPUTER" is used to denote such class of system that can advance

knowledge and generate insight that would not be otherwise be possible or that could not be captured in time to be actionable. They are the indispensable tools for solving the most challenging and complex scientific and engineering problems including the simulation and modeling of physical phenomena. As the technology advances to a new era the core component of any computation i.e. data is not stable and to compete with such petabytes of data's modest computing system is required. However many fast processing systems were developed, yet they are dawdled by scaling, timeliness, architectural design and ability to address important issues.

The advancement in petascale computing technologies that will overcome the processing and performance constrain of computing resources. One motivation of such computing is to aggregate the power of multiple system in to single system to study the high end calculation intensive task that cannot be possible with single core system. To achieve this goal, proper understanding of system tools, software and underlying hardware is essential.

The development can be traced back to 1960 with the initiation of first supercomputer CDC 6600 by Seymour Roger Cray at Control Data Corporation. With time their developed more such system that can exploit the processing speed and performance, few with thousands of processors and others more than that. Though the market was flourished by European design yet, India was no way out from the challenge, and the development was marked by India's first supercomputer Param 8000 built in 1990 by Center for Development of Advanced Computing (CDAC).

3.1 HPC for climate studies

The Earth is good as a single interlinked and self-regulating system. It's subsystems, like atmosphere, hydrosphere, cryosphere, geosphere and biosphere purpose together and their interactions are significant and complex. The energy and material transport within and across subsystems happen from local to global scale in varying space and time. Improved and reliable forecast of weather and climate involves integration of observations using very high resolution dynamical models with realistic illustration of all physical processes and their complex nonlinear interactions. Since weather is an initial value problem, accurateness of the initial condition is as important as the accuracy of the model. Thus, data assimilation is a key component of weather predictions. As conventional data coverage is spatially and temporally limited, satellite data provides better coverage in both space and time. About 90% of the data that goes into the assimilation of any analysis-forecast system consist of data from satellite and rest from in situ platforms. In addition, it is important that adequate computing facility is available for carrying out various numerical experiments pertaining to various programs. This includes augmenting the computational power for the training school where hand on training are to be conducted with high resolution state of the art weather and climate numerical models, conducting research and development work for improving forecasts in the short, medium and long range scales for monsoon mission programs that involve sensitivity experiments for various physical processes, the impact studies of different physical parameterization schemes etc., data impact studies, ensemble prediction models with more members, climate change scenario generation for hundreds of years etc. [4]. Additional to this, it is essential to carry out studies related to observation simulation experiments (OSE), observation system simulation experiments (OSSE) and targeted observation experiments that can guide the planners on the location and type of observations that are crucial for the numerical models. Accordingly observation network can be better formulated. This is highly compute intensive job. Large number of numerical experiments shall have to be carried out to identify these crucial locations where observation network need to be strengthened. Therefore, it is seen that the entire range of research work involves simulation runs of multiple versions of the same high resolution analysis forecast model which means the utilization of HPC time as well as storage also becomes manifold (directly depending on the total number of experiments undertaken by each student). In order to study the effect/impact on a large temporal scale (from monthly to decadal to 100s of years), these runs are to be undertaken accordingly. In addition, for understanding the micro scale process studies one has to go for extremely high resolution models that can resolve scales of the cloud and related processes. Thus these entire ranges of studies require not only high level of computer storage, high computational power as well.

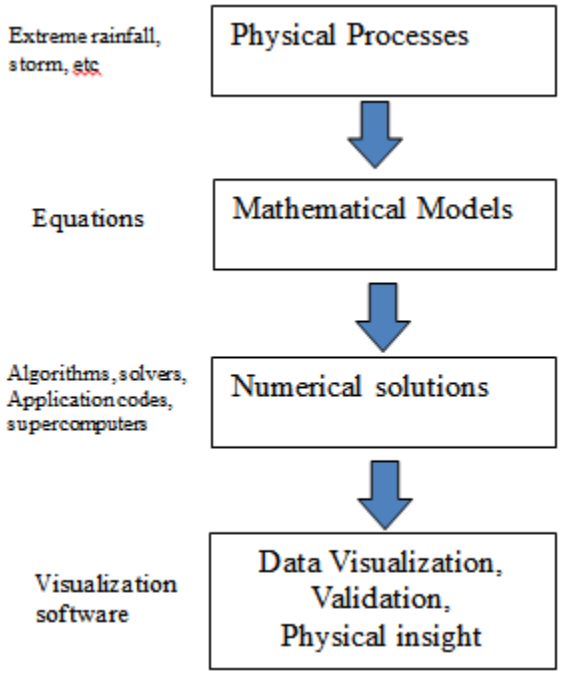

Fig 1: Flow diagram of climate model simulation and Visualization in HPC Platform

3. CLIMATE CHANGE STUDIES IN HIGH PERFORMANCE COMPUTING

This commodity approach to high-performance computing especially critical to climate change studies is

because of Scalable and Multicore Computing Strategist [3]:

First, one must simulate hundreds to thousands of Earth years to validate models and to assess long-term consequences. This is practical only if one can simulate a year of climate in at most a few hours of elapsed time at reasonable cost. Each of these simulations must be of sufficient fidelity (i.e., temporal and spatial resolution) to capture salient features. Today, for example, most climate models that are run for several hundred to several thousand simulated years do not explicitly resolve important regional features like a hurricane. These are large-scale, capability computing problems (i.e., ones requiring the most powerful computing systems).

Second, to understand the effects of environmental changes and to validate climate models, one must conduct parameter studies (e.g., to assess compassion to different conditions such as the rate of CO2 emissions or changes in the planet's albedo — its reflectivity and solar energy absorption). Each of these studies involves hundreds to thousands of individual simulations. This is only practical if each simulation in the ensemble takes a modest amount of time. These are large-scale, capacity computing problems (i.e., ones requiring ongoing access to multiple, large-scale computing systems).

Third, understanding the sensitivity of physical and biogeochemical processes to social, behavioral and economic policies requires evaluation of statistical ensembles and many model variants. These are hypothesis-driven computational scenarios that are only possible after the physical and biogeochemical processes are understood, requiring additional capacity and capability computing.

High-performance computing expertise permits to efficiently achieve high-performance throughputs for intensive CPU load applications. The scale of climate data (observed and model-simulated) is growing rapidly. Besides scalability there are many additional challenges like data-related challenges including complexity, high dimensionality, and the spatio-temporal nature of the data; and hardware-related challenges arising from increasing complexity and capabilities of new HPC systems. We are developing computational methods and tools that leverage the latest technologies (such as multicore architectures, GPGPU accelerators, etc.) to enable the analysis of large-scale datasets.

HPC technology generally licenses to efficiently achieve high-performance throughputs for intensive CPU load applications.

4. REGIONAL CLIMATE MODEL

The main requirement is to understand the model. Climate models help us understand the physical processes that govern climate and enable us to predict climate changes. They range from conceptual models to models of intermediate complexity to comprehensive 3D models with sophisticated representations of the major components of the Earth System.

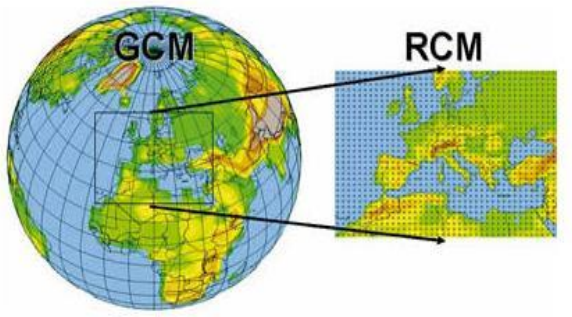

Fig 2 : Resolution difference between GCM and RCM

The 3D models typically quantify the interactions of the atmosphere, oceans, land surface, and ice. Can be used to simulate a variety of processes and feedbacks (e.g. climate change) [6].

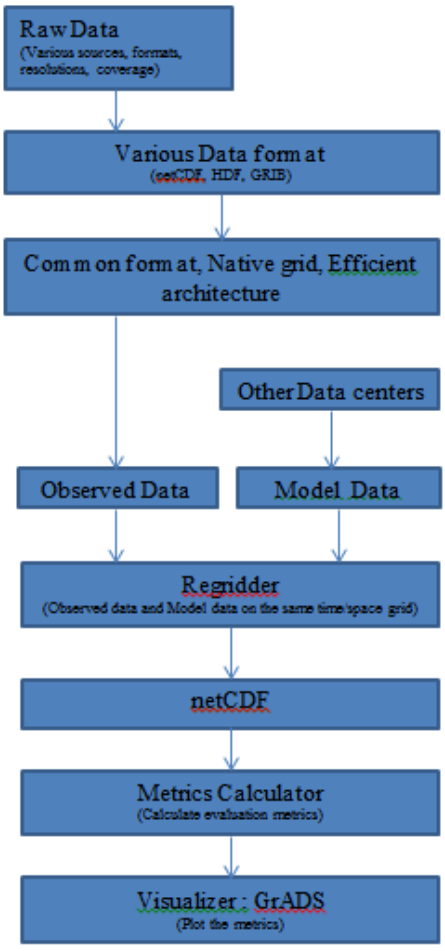

Fig 3 : Flow chart of Regional Climate Model

For viable regional climate change experiments three conditions must be fulfilled: (a) The GCM climate must be realistic over the region of interest (b) Model resolution has to be high enough to capture the basic mesoscale forcings of relevance and its domain large enough to allow the full development of mesoscale circulations and (c) Model domain has to be large enough to prevent the coarse-scale lateral boundary conditions from dominating the solution over the RCM area, but small enough to prevent its

circulation from departing far from that of the driving GCM, otherwise physical consistency between the GCM solution is not maintained.

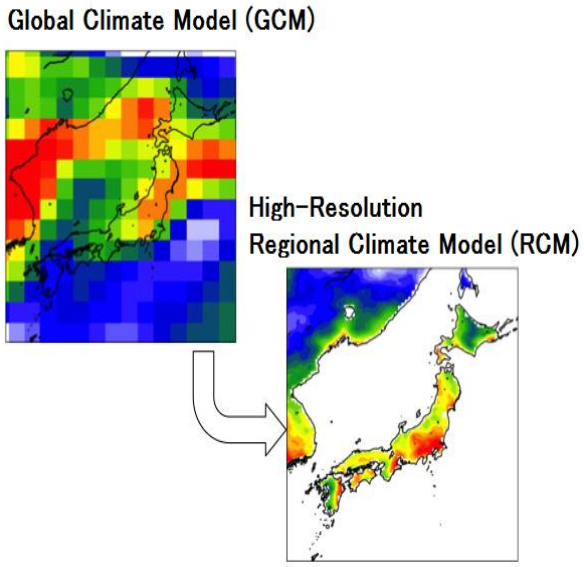

Fig 4 : High Resolution RCM

Steps to run RegCM
Step-1: Make a folder outside from your main RegCM folder e.g. RCMSimulation
Step-2: In RCMSimulation folder link Bin from your main RegCM folder
Step-3: Copy test.in from RegCM to your RCMSimulation folder
Step-4: Put your data in RCMSimulation folder
Step-5: Make two folder INPUT and OUTPUT in your RCMSimulation folder

Step-6: Define your number of grid point in your test.in
Step-7: Once you done all the above steps then prepare your initial and boundry conditions by doing:
(a) ./Bin/terrain test.in
(b) ./Bin/SST test.in
(c) ./Bin/icbc test.in
Step -8: Once you made your ICBC then just give the run by :
mpirun -np 1 ./Bin/regcmMPI regcm.in > RCM.output

5. VISUALIZATION SOFTWARE AND DATA FORMAT

5.1 Visualization Software (GrADS)

The tool used to display large climate data is the Grid Analysis and Display System [7] (GrADS). It is an interactive desktop tool used for easy access, manipulation, and visualization of earth science data. In GrADS there are two data models for handling gridded and station data. GrADS is freely available over the internet and it is implemented worldwide on a variety of commonly used operating systems.

5-Dimensional data environment is used by GrADS and the four conventional dimensions are longitude, latitude, vertical level and time. An optional 5th dimension for grids is generally implemented but designed to be used for ensembles. By the use of data descriptor file the data sets are placed within 5-D space. GrADS handles grids like regular, non- linearly spaced, gaussian, or of variable resolution. With correct spatial and time registration, data that are from different data sets may be graphically overlaid. An operations can be executed interactively by entering FORTRAN-like expressions at the command line. GrADS provides a rich set of built-in functions, but users may also can add their own functions as external routines written in any programming language.

Data may be displayed using different graphical techniques such as line and bar graphs, smoothed contours, shaded contours, scatter plots, wind vectors, grid boxes, streamlines, station model plots, and shaded grid boxes. Graphics output may be postscript or an image formats. The user as the option to control all aspects of graphics output even though GrADS provides geophysically intuitive defaults. GrADS has a programmable interface that allows for sophisticated analysis and display applications.

6.1 WEATHER PREDICTION

The various methods used in prediction of weather are explained in some papers [11], i.e. Synoptic Weather Prediction: It is the traditional approach in weather prediction. Synoptic refers to the observation of different weather elements within the specific time of observation. In order to keep track of the changing weather, a meteorological center prepares a series of synoptic charts every day, which forms the very basic of weather forecasts. It involves huge collection and analysis of observational data obtained from thousands of weather stations.

Numerical Weather Prediction: It uses the power of computer to predict the weather. Complex computer programs are run on supercomputers and provide predictions on many atmospheric parameters. One flaw is that the equations used are not precise. If the initial stage of the weather is not completely known, the prediction will not be entirely accurate.

Statistical weather prediction: They are used along with the numerical methods. It uses the past records of weather data on the assumption that future will be a repetition of past weather. The main aim is to find out those aspects of weather that are good indicators of the future events. Only the overall weather can be predicted in this way.

7. CONCLUSION

As climate data is huge we need to implement data and model in High performance computing environment by using data mining approaches. A data mining toolkit will be on place to which will be meant specifically for large-scale climate data analysis to understand the climate change over India with a capability of spatio-temporal analysis at various scales.

REFERENCES

[1] Intergovernmental Panel on Climate Change. "Climate Change 2007: Fourth Assessment Report (AR4)" (2007).
[2] Tollefson, J. "Climate war games", Nature, Published online 5 August 2008, doi: 10.1038/454673a.
[3] Dan Reed. "High Performance Computing: Enabling Climate Change Analysis", Scalable and Multicore Computing Strategist(2011)
[4] Ministry of Earth Sciences. "High Performance computing System". Published on 4 June 2015.
[5] The Hans India. "Supercomputing Mission". Published on 13 May 2015.
[6] L.O.Mearns, F.Giorgi, P.Whetton, D.Pabon, M.Hulme, M.Lal. "Guidelines for Use of Climate scenarios Developed from Regional Climate Model Experiments", Published on 30 October 2003
[7] George Mason University. "Documentation of GrADs" in Centre for Ocean-Land-Atmosphere Studies, Institute of Global Environment and Society.
[8] A K Soniyapriyadharshini and Dr. P B Ramesh babu. "High-through functions and classification of Data Mining Inter-specific with Web Mining" 0975-8585 (2013)
[9] Mrs. Jasmeet kaur and Mrs. Jigisha Pandya. "Knowledge discovery in Databases" in IBMRD's Journal of Management and Research, print ISSN: 2277-7830 (2013).
[10] Sarah N. Kohail, Alaa M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis", IJICT Journal Volume 1 No. 3, 2011
[11] S. Kotsiantis and, "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values", World Academy of Science, Engineering and Technology, 450-454, 2007