A PANORAMA OF SCIENTIFIC DATA MANAGEMENT IN WATER AND CLIMATE MODELING

S. PUROHIT, A. KESARKAR and A. KAGINALKAR
Centre for Development of Advanced Computing, Pune University Campus, Pune

ABSTRACT *Improvements in the mathematical model performance of weather and climate have been complemented by the development of supercomputing infrastructure and technological advancement in scientific data management, data processing and data visualizations. This has enabled meteorological community to make significant progress in the areas of operational weather forecasting as well as in research and development activities. The new era in Atmospheric Sciences viz. ensembling of parameters obtained from the model output for quantitative weather and climate predictions and research; of high resolution downscaling of the atmosphere for the prediction of socio-economic impacts on various spatial and temporal scales; long term integration of mathematical models for climate change predictions and their possible impacts on the atmosphere of next decades, require huge computational resources as well as effective scientific data management. This paper provides a panorama of some of the recent developments in the scientific data management for weather and climate modeling. The paper discusses the key requirements for scientific data management weather community and proposes a distributed multicomponent environment for inter-disciplinary scientific data management (C-DME).*

Key words Climate data; supercomputing; data management; climate modeling; data communication.

INTRODUCTION

The rapid advent in the techno-scientific research related to high resolution downscaling/ forecasting of the weather in the last few decades has led to the development of complex numerical models. These models are developed for several spatio-temporal scales. The different studies pertaining to the use of these models in the seasonal weather forecasting (using Global Climate Models), in the regional weather forecasting (using Mesoscale Numerical Models), in the air quality modeling (using statistical-dynamical models for air quality forecasting) and in the climate change forecasting (using Coupled Ocean - Atmosphere Global Climate Models) are useful in the planning and preparedness, identification of vulnerable situations and for mitigating the risks associated with weather on different spatial and temporal scales.

Such research has direct implications in the decision making and human operations and, hence, ensuring the timely availability of value added, customized data for conducting this research becomes the prime responsibility of a researcher/ data provider. An estimate of data storage requirements for weather and climate research can be obtained by looking at the World Data Centre (WDC) database of Model and Data group (M&D, 2007) of Max Planck Institute (MPI), which has been rated as world's largest database. This database contains approximately 111 Terabytes (TB) data from model simulations and approximately 220 TB data related to the information on climate research. An additional 6 Petabytes (PB) of

information is stored in their magnetic data tapes. This data will be used by the Intergovernmental Panel for Climate Change (IPCC) for preparing a report on Climate Change which will be published by the end of 2007. To generate such a vast database requires enormous multi-disciplinary efforts. For example, the dataset required for atmospheric models as a input might require recording and dissemination from weather observatories, aircrafts, weather radars, flying balloons, ships, buoys and satellites. The order of the number of such global observations is approximately 10^6 records per synoptic hour (every 3 hours) on real-time basis. Thus, it poses a daunting challenge for researchers to manage such wealth of information. The critical problem faced by the research community is the unavailability of software to manage such huge data, which has to be accessed through different network methodologies and is stored in various formats, in real-time. The data acquired from different locations can be assimilated to form a gridded dataset that is useful as an input to global General Circulation Models (GCM). The outputs obtained from GCM are used to generate initial and boundary conditions required for initialization of the mesoscale numerical models. The outputs obtained from mesoscale models are used to generate the regional weather forecasts.

Hence, the real-time data acquisition methodologies play an important role in data interconversion and interoperability between different formats. The raw data in the form of model outputs are generally available from atmospheric centres with supercomputing/ high performance computing infrastructure. Over the last decade, the developments in grid computing have added new dimensions to the model data availability, computing, distribution and management. Moreover, the developments in the field of scientific data visualization and availability of web technology for data dissemination have had a significant impact on the scientific data management for weather and climate.

This paper presents a comprehensive panorama of recent developments in the technologies related to scientific data management for weather and climate research and operations. The next section of this paper discusses about the issues in scientific data management related to weather and climate modeling. In subsequent sections the role of high performance/ supercomputing/ grid computing in scientific data management is described and the inter-disciplinary Data Management Environment for atmospheric research (C-DME) proposed by Centre for Development of Advanced Computing, (C-DAC) India is introduced.

ISSUES IN THE SCIENTIFIC DATA MANAGEMENT FOR WEATHER AND CLIMATE MODELING

The need of scientific data management is frequently expressed by scientists working in the field of high-resolution weather and climate simulations trying to get the answers to many questions related to the behavior of atmospheric parameters of the present date, of the next decade and of the coming centuries. These studies demand high resolution modeling of weather and climate to understand the consequences associated with changes occurring in atmosphere as well as oceans. The increase in model resolution leads to an increase in the number of grid points used for horizontal spatial differencing of complex model equations. For example, consider a GCM that uses global Gaussian grids and a spectral triangulation method

for horizontal discretization. Table 1 shows the increase in the number of grid points per variable with the increase in the horizontal resolution of this model.

Table 1 Horizontal spatial differencing: spectral triangular (global Gaussian grid).

Model resolution	Number of points on Equator	Spatial resolution	Vertical levels
T80 L18	256 X 128	~ 1.4 ° X 1.4 ° : 130 X 130 Km	18
T126 L28	256 X 128	~ 1.0 ° X 1.0 ° : 130 X 130 Km	28
T170 L42	512 X 256	~ 0.7 ° X 0.7 ° : 78 X 78 Km	42
T254 L42	768 X 384	~ 0.7 ° X 0.7 ° : 50 X 50 Km	42
T382 L64	1152 X 576	~ 0.3 ° X 0.3 ° : 40 X 40 Km	64
T511 L64	1536 X 768	~ 0.23 ° X 0.23 °: 26 X 26 Km	64
T799 L91	2048 X 1024	~ 0.17 ° X 0.17 °: 19 X 19 Km	91

This model is found to be useful for simulating complex coupled planetary systems such as Indian Summer Monsoon (ISM). Ratnam et al. (2007) in their recent studies, simulated the ISM of year 2005 using this model with resolution of T170L42 and demonstrated that the increase in the resolution leads to realistic simulation of spatio-temporal and intra-seasonal features associated with ISM as compared to that simulated by the model with coarser resolutions. The simulated total seasonal ISM rainfall for the year 2005 is as shown in the Fig. 1.

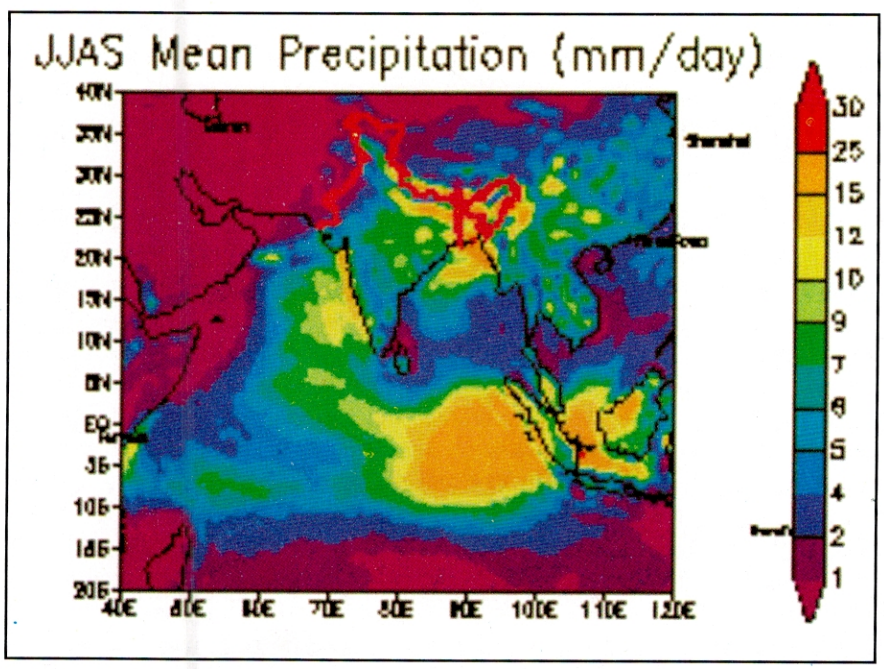

Fig. 1 Simulated seasonal ISM rainfall from GCM with T170L42 resolution for 2005.

It can be observed that the ISM rainfall is realistically simulated over the regions of the elevated orography such as Western Ghats, Northwest India and also over the low precipitation regions such as Northwest Gujarat and Rajasthan. Such efforts are useful for Long Range Forecasting (LRF) of ISM and important for socio-economic policy making. As seen from Table 1, for the GCM resolution of T80L18, the total number of computational grid points per model variable are 304200 while those for T799L91 are 190840832. It can be seen that for each variable there is an increase of approximately 628 times in the requirement of resources for the execution of the model T799L91 as compared to T80L18. This in turn increases the per variable data storage requirement of T799L91 by at least 5024 times as compared to T80L18 if we consider each variable is of a size of 8 bytes (64 bit).

Thus, if the disk space required to store a single variable of this model having resolution of T80L18 is 1 MB then the same variable will require disk space of about 5.024 GB if we simulate the weather using a resolution of T799L91. Such atmospheric/ oceanic models store a few hundred variables. Therefore, the increase in the resolution of the model causes a multi-dimensional increase in the data storage requirements and imposes a stress on the efficiency and effectiveness of the data management system. There are many issues and requirements related to scientific data management for weather and climate. Some of them are addressed below:

Inter-Disciplinary Studies and Data Requirements

Weather and climate of a region gets affected by various natural as well as anthropogenic factors and has important socio-economic consequences. It is believed by the scientific community that the changes in the atmospheric chemical cycle, alteration of land use fractions and practices, vegetation fraction, changes in the sensible heat due to anthropogenic activities has altered the weather and climate of the globe. These studies are multi-disciplinary in nature and deal with geographical, geological, ecological, socio-economic and meteorological information, simultaneously.

Figure 2 shows the global land use fractions used in the GCM. This dataset is generally updated once in 10 years. However, the vegetation fraction datasets as shown in the Fig. 3 need to be updated every fortnight for the realistic simulations of the atmosphere. Thus, the regular availability of updated datasets is also a major concern for realistic modeling of weather and climate. On the other hand, some of the studies related to the operational weather forecasting require the background information of the weather over the region in the form of climatology and variability of several atmospheric parameters. Up-to-date information on such parameters helps weather forecasters to estimate the severity of the occurring weather, speculate socio-economic consequences associated with it and inform different users about the severity of the event. Such information as a database if available to the forecasters in terms of probabilities and anomalies of selected weather parameters, facilitates the visualization of the general behavior of the atmospheric system over that region. To judge the similarity between evolving atmosphere and previous occasions, similarity/ pattern recognition is an essential component.

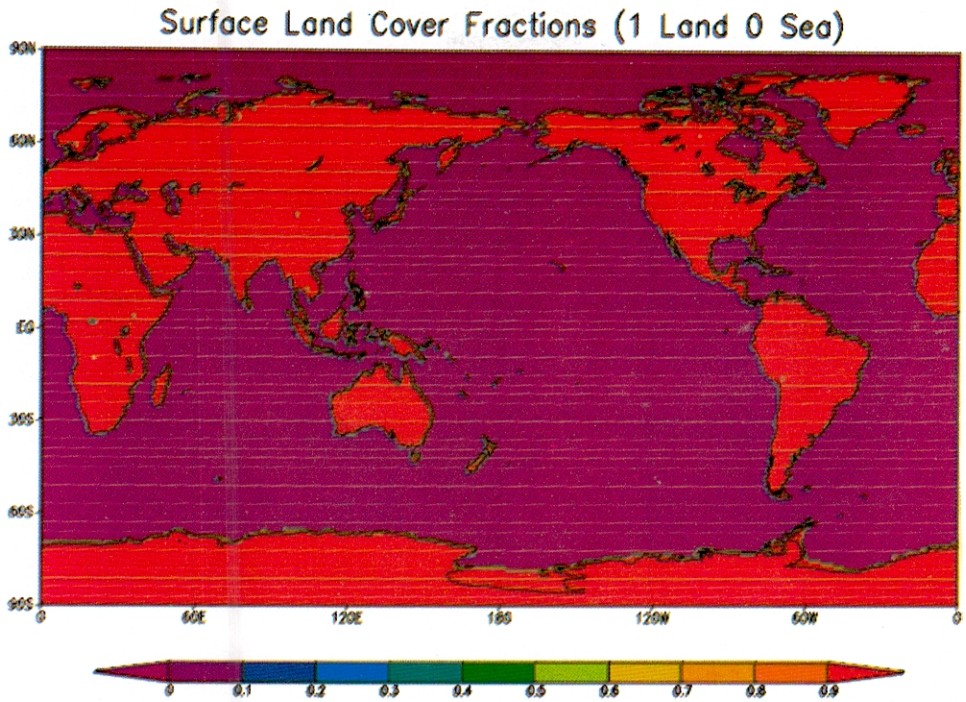

Fig. 2 Global surface land fractions.

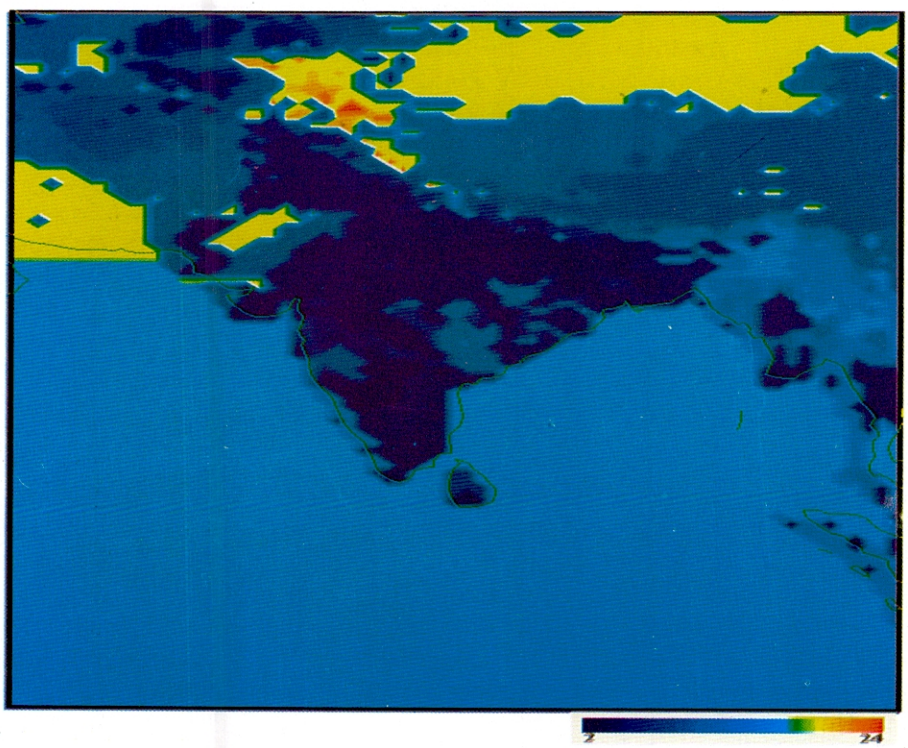

Fig. 3 Vegetation fraction over Indian subcontinent.

If such similarity is observed in previous databases then it strengthens the forecaster's confidence on the issued weather forecast. These multi-disciplinary requirements demand for an integrated, comprehensive robust framework of data mining and knowledge discovery system for weather and climate research and operations.

Different Data Communication Methodologies

Even though the data processing capacity has demonstrated tremendous growth in the last decade, the data distribution capacity has not grown to the desired extent and there is always a demand for reliable, meaningful data transmission in an effective manner. Given this ever increasing demand for data communication in future, either in real-time mode to perform experiments or in delayed mode to create decision support methodologies, there is a need of the effective and efficient data communication, which incorporates the best available technology of the date. The various technologies presently used for communication of weather data are listed below.

Global Telecommunication System

Global Telecommunication System of World Meteorological Organization (WMO) is an infrastructure used to collect and disseminate meteorological data. It is an ensemble of integrated networks which interconnect meteorological centers across the world. It has a hierarchical structure of national, regional and global telecommunication links and uses different telegraphic media including satellites and the internet. It is used to transmit real-time high priority data and non real time archives in ASCII (SYNOP, METAR, TAF, TEMP, PILOT, etc.) as well as binary (GRIB, BUFR) formats. This service can be employed for low resolution meteorological image transmission but is less effective as compared to the recent meteorological data transmission methodologies. This infrastructure has well defined procedures and data sharing methodologies, but use of proprietary high level protocols is not supported. There is also a restriction on the volume of the transmitted data, unavailability of infrastructure which facilitates the distribution of weather software / codes to process this data and it lacks support to provide adhoc access to this system.

Internet

Internet has become a popular medium for the communication of the weather data in the recent times. Internet is used for ubiquitous mining of weather data from the distributed sources in real-time. A large number of distributed sources of data available on the Internet can be employed for its fruitful utilization. This protocol provides an access to the data in different formats and is highly efficient. The software required for decoding of this data can also be distributed through this protocol. This protocol is widely used for distribution of near real-time computing of weather parameters across the heterogeneous network and provides cost effective solutions. However, this protocol is always limited by the availability of network bandwidth and questionable data security.

E-mails

E-mails use Hypertext Transfer Protocol (http) for data communication (W3C, 2007). This is one of the most efficient methods for the delivery of the weather data as it delivers weather data to specified authentic users with in few seconds across the globe. However, its main limitation is that large volume of the data cannot be transferred using e-mails.

Satellite Communication

Satellite communication is a very useful medium for inter-continental weather data exchange especially in the GTS data transmission. The satellite cloud pictures available from the geostationary and polar meteorological satellites, provide a snap shot of different meteorological parameters. There are satellites such as Tropical Rainfall Measuring Mission Satellites with onboard precipitation radar for monitoring of rain processes in the atmosphere. Due to availability of satellites for weather monitoring purposes, the weather forecasting practices have undergone a major revolution. There are many satellites used for monitoring ocean winds and sea surface temperatures. WMO's World Area Forecast System (WAFS) has been operational since 1996 based on satellite communication. WAFS products include: Wind and Temperature Forecasts spanning from 6 to 36 hours - updated at least twice daily. This system is useful for dissemination of weather data to and from remote locations where GTS links are difficult to manage.

Geographical Positioning System (GPS)

GPS has been mainly developed for global navigation purpose and uses a network of satellites to locate global positions accurately. It has been observed that the signal loss in the GPS link is proportional to the vertical distribution of the moisture. This property is made use of for profiling the vertical distribution of the moisture, temperature, pressure with altitude. This data is highly useful for assimilation in weather models.

Variety of Data Formats

There are various meteorological data formats which can be used for data storage and distribution, varying from character format, packed binary, to self-describing formats which have evolved over the duration of the last few decades. The most commonly used meteorological data formats are described below.

Conventional WMO Formats

Conventionally, meteorological data is transferred in terms of coded weather messages to ensure the efficient transmission of weather information over GTS. Mainly, weather messages such as Surface Weather Message (FM 12-VII SYNOP), Upper Air Radio Sonde Raw Winds observations (FM 35-X Ext. TEMP, FM 36-X Ext. TEMP SHIP and FM 38-X Ext. TEMP MOBIL), Pilot Balloon Observations

(FM 32 PILOT/FM 33 PILOT SHIP, FM 34 PILOT MOBIL), Routine Aviation Weather Report (FM 15-V METAR) and Terminal Aviation Forecasts (ICAO, Annex 3) are recorded and transmitted through GTS. These formats are relatively complicated for decoding as compared to the recent weather data formats as they are not self describing formats.

Hierarchical Data Format

The Hierarchical Data Format (HDF, 2007) is a multi-object file format useful for storing and transferring scientific data. A large number of datasets are currently stored in HDF because of its stability and flexibility. With the rapidly increasing volumes of data in HDF, scientists are facing problems of effective and efficient data access. The recent Metadata technology developed for HDF provides a simplified access to this data, which ensures its long-term usability.

netCDF: Network Common Data Format

NetCDF (NetCDF, 2007) is a set of interfaces for array-oriented data access. The netCDF libraries support a machine-independent format for representing the scientific data. Together, the interfaces, libraries, and format support the creation, access, and sharing of scientific data. This format is widely used for storing data related to atmospheric and climate models. This data format is portable, self-describing, direct access, appendable, easily sharable and achievable.

BUFR: Binary Universal Form for the Representation of Meteorological Data

The WMO code form FM 94 BUFR is a code designed to represent a continuous binary stream of any meteorological data. It is a self-descriptive format. This data format has the strength of accommodating the changes in its structure. Because of data descriptors used for describing this data, any missing data does not get coded and transmitted along with this data. Another advantage of the code is that it is a character oriented code. It is relatively easy to decode and accelerates conversion of a message into a machine useful numeric format. Further, this format has been developed to accommodate present and future observational practices.

GRIB: Gridded Binary Data Format

It is an efficient data format used for transmission of the large volumes of gridded data to automated weather centers over high-speed telecommunication lines using modern protocols. The GRIB codes are packed to make them more compact than character oriented bulletins, which result in faster transmissions. This format can equally well serve as a data storage format, generating the same efficiencies relative to information storage and retrieval devices.

Earth Science – eXtensible Markup Language (XML)

The extensible markup language is nowadays becoming a popular standard for exchanging data over the web. This is a platform independent self-describing format

found to be useful for loosely coupled data exchange and for flexible semi-structured data.

Data Archival Practices Used by Different Data Providers

In meteorological modeling, the model outputs are archived and distributed in the different coded message formats depending on the model that generates them. This heterogeneity limits the usefulness of such data and creates data-application and interoperability problems for development of scientific tools. There are efforts to eliminate these limitations by providing global standards for information interchange and coding. However, the coordination of different data providing agencies with the different categories of the data users is a major limitation for the effective use of datasets. The continuous evolution of data coding standards as well as archiving methods with the development of new technologies also remains one of the major concerns for interoperability. Thus, any data management system developed for this purpose should be compatible with the latest methodologies for easy updation of data decompression and decoding services.

Data Interoperability

Data interoperability eliminates the barriers associated with data sharing. It facilitates effective data access, transformation and dissemination across the network or computational grid interfaces. The Earth Science Markup Language (ESML) is an elegant solution for data interoperability problem. It is an interchange technology that enables data (both structural and semantic) interoperability between applications without enforcing a standard format mentioned earlier. ESML uses a schema to describe the structure of the data file. The applications accessing weather data available in different formats can be developed easily using the ESML library. The updates to data coding/ decoding standards and archiving methodologies can be easily incorporated by modifying the schemas.

ROLE OF HIGH PERFORMANCE/ SUPER/ GRID COMPUTING

A rapid proliferation of supercomputing technologies has been witnessed in the last decade due to developments in the cache-based microprocessors which are useful for building high-end computing facilities. The architecture of these microprocessors has generality, scalability and cost effectiveness. The architectural design of super-computers which offers a balance between memory performance, network capability and execution rate with the help of these microprocessors has also improved over the last decade. Because of these developments in supercomputing, at present, various international organizations involved in the field of weather and climate have Tera-scale computing infrastructure available for research and development as well as operational purpose. These facilities have been developed over a number of years and have diverse software and hardware artifacts. For example, National Centers for Environmental Prediction (NCEP) of USA uses Global Forecast Model T382L64 (resolution of 35 km on equator) for operational weather forecasting. The computing power available for this purpose is 7.8

Teraflops (TF). The other national agencies of USA such as Forecast Systems Laboratory (FSL) has a 6.7 TF system dedicated for operational weather forecasting using Weather Research and Forecasting Model (WRF). Geophysical and Fluid Dynamics Laboratory (GFDL) of USA uses a 3.0 TF machine for operational weather forecast using the R30 Global Coupled Model. National Centre for Atmospheric Research (NCAR), USA, uses an 8.3 TF machine for conducting research and development using various climate models such as CCSM, CAM and mesoscale models such as MM5 and WRF.

The Japanese Agency for Marine Earth Science and Technology (JAMSTEC) has Earth Simulator, for research related to oceanography and climatology. Earth simulator has a peak power performance of 40.0 TF and is associated with an online storage of 250 TB and secondary storage of 1.5 PB. European Centre for Medium Range Weather Forecasting (ECMWF), located in United Kingdom also has operational responsibility of weather forecasting for the European region and has supercomputing facility with a peak performance of 16 TF available for this purpose. The quantum jump in the super computing facilities available at NASA provided unprecedented opportunities to NASA scientists to develop a Finite Volume Global Mesoscale model with the resolution of 0.125° for operational forecasting and research. This model is used specially for operational hurricane forecast to minimize the hurricane related damages.

This trend in development of high performance computing systems has also shown an impact on the weather and climate research and operational weather forecasting practices. Due to availability of supercomputing infrastructure, operational forecasters and researchers can now plan for weather simulations at cloud resolution scales. Similarly, climatologists can now plan simulations for the next centuries to study the impact of human endeavors on weather and climate. The weather/ climate forecasts generated from these simulations are in the form of high volume datasets. Therefore, these developments have in turn increased the need for efficient data and storage management.

The storage management systems need to cater to large datasets generated over different computing platforms, stored in different formats and archived using different methodologies. Most of the decision makers or targeted users (researchers, academicians, etc.) do not have in house expertise for handling such datasets and extracting useful information from the data. The development of sophisticated, customized, data processing, visualization, data-mining and knowledge extraction tools have helped decision makers, meteorologists and climatologists in deriving scientific understanding of weather and climate parameters from these datasets. However, a majority of such efforts do not provide an integrated solution that can cater to the different needs as described above.

There is, thus a need for development of software infrastructures which create datasets, maintain them, evolve them with time, federate them into active digital libraries of scientific data and also facilitate the user for their efficient usage. Development of such infrastructures requires multidisciplinary efforts pertaining to heterogeneous data management. The challenge is to design a data management system that can anticipate the needs of the user community for the next 5 to 10 years.

C-DAC'S INTER-DISCIPLINARY DATA MANAGEMENT ENVIRIONMENT (C-DME)

There are various challenges in the development of a new ordered system for weather and climate data management. The main challenge is the transformation of data into an integrated resource that supports both present and future requirements. It is also a great challenge to address some of the issues related to data management such as data awareness, understanding, variability, redundancy, access, cataloging, data mining, knowledge discovery etc. This section discussed an approach for the development of an inter-disciplinary data management environment especially applicable for atmospheric science research and catering to the several challenges discussed above.

Adaptation of Global Standards and New Methodologies of Meteorological Data Formats

Platform Independence

The meteorological community always performs computations on different computing platforms that have different computing power, numeric precision, different procedures to store the information, and heterogeneity at the hardware level. It is very difficult to maintain the same operating environment and computing configuration in different geo-distributed meteorological stations/ research laboratories. If this data needs to be exchanged, then significant human intervention is required to manage the data inter conversion and compatibility issues. Therefore, there is a great need for a machine independent interface between system level software and meteorological data acquisition and analysis applications. New generation (NetCDF, HDF) data types described earlier inherently incorporate platform independence to a large extent. Platform independence will be one of the major design requirements of the proposed C-DME.

Another important aspect of the storage of data is data precision. As multi-disciplinary data from different computational systems is of different precisions, a uniform 128 bit precision is proposed for the common data structure in C-DME.

Self Describing Formats

Generally there are many attributes or metadata (such as name of the variable, dimensions, datatype, unit of measurement, accuracy of measurements, data location, data source, instrument used for measurements, etc.) associated with different meteorological/ climatological datasets. These attributes may vary based on the locations, recording instrument, objectives of the experiments, type of utility used for generating this data, etc. The new data types and data standards provide an abstraction to such datasets making it self-describing. A self -describing data format contains structural metadata, which is used by a corresponding runtime library to navigate through the file and improve I/O performance by allowing for direct access to a particular dataset within a file or to specific parts of a dataset. The provision of such libraries in the C-DME will make it more flexible to handle different datasets and perform various operations over these datasets.

Components Approach for Development of C-DME

For more than a decade, good software development practices have been based on a divide and conquer approach in software architecture design and implementation (Brown and Wallnau, 1996). In such approach, a software system is decomposed into manageable components or units to provide maximum cohesion within a component and minimizing the coupling between different components (Parnas, 1972). The renaissance in the component based approach for software development that occurred in the recent decade is based on the advances in the object oriented approaches and the economic viability of large softwares. The object oriented approach is based on the development of an application system through the extension of existing libraries of self-contained operating units, thus ensuring the economic viability of development of large software by making maximum use of existing software and reducing the need for development of new codes. This approach demands the segregation of off-the-shelf components; components having significant aggregate functionality and complexity; self-contained components; and components that can be integrated with other components to achieve required system functionality. The use of off-the-shelf components is a major challenge, since these components have varied pedigree, many unknown attributes and are of varied quality. The meteorological data format libraries can be treated as off-the-shelf components as they have their own format for describing the same meteorological dataset along with different protocols for their storage.

Even though this approach of using the off-the-shelf components is based on the reusability of the developed libraries, the enabling of common functionality to each of the component libraries describing different meteorological formats may pose a challenge for their integration and also for component maintainability. The interface that provides the access to component functionalities such as data encoding/ decoding, exploration, interpretation, interconversion and mining to the user needs to have good performance, reliability and reusability. Therefore, the unknown internal operations associated with the execution of off-the-shelf components need to be analyzed through hands on evaluation. Further, the designing of self contained components by partitioning the total required functionalities of such large database systems into components jointly achieving the required functionality is a major task and may require redesigning of the some of the off-the-shelf components. The coordination model designed at abstract levels for describing interaction of different components to carry out the end user functionalities needs to be validated and enacted under heterogeneous lower level infrastructures. This design requires plug and play in such a way that, components designed to interact with each other can have ease of evolution at unit level and can be repaired or updated or swapped as a plug-replaceable unit. The constraint of time for development, upgradation and repair of the system demands for innovative methodologies to support the task of technology insertion. The simplex architectural designing methodology (Sha et al., 1995) based on open system components is developed to address some of these concerns. It involves designing of layered software which allows reliable and safe upgrade of an online system in realtime and also supports the use of analytical redundancy through fault tolerance and self healing. The architecture of C-DME presented in the Fig. 4 is based on aforementioned requirements and design features.

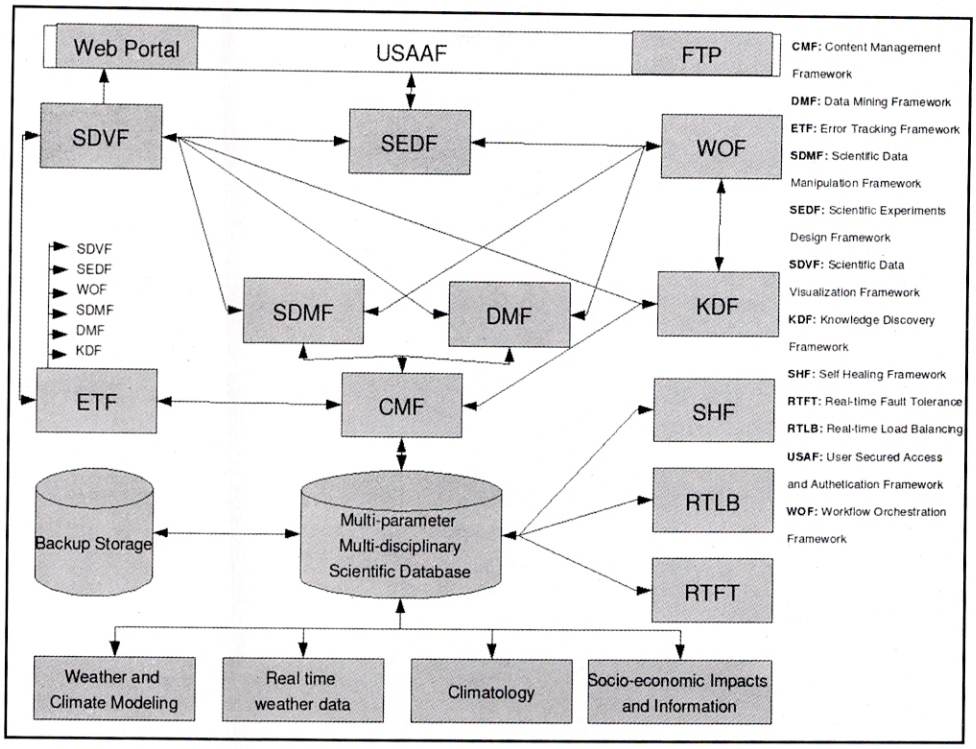

Fig. 4 Architecture of integrated scientific data management environment for atmospheric research.

Components of C-DME

The C-DME architecture has been designed in such a way that all the components are subject-oriented and robust. The independent components can be upgraded, extended, expanded and enhanced in their infrastructural setup without affecting the operations of other components. This architecture is service-oriented for real world scientific data management. The architecture is designed to adopt global standards and all the recent methodologies used for meteorological data storage and archival. The detailed functionality of each component is described below.

Secured User Access and Authentication Framework (USAAF)

The efficiency and easy access of the weather/ climate information is vital because of the short time period available to generate the weather forecast as well as to carryout the research. To avoid delays in data access and data processing, secured and authenticated access to the information is critical. This avoids the unprecedented load on the system as well as avoids congestion in the network traffic. Further, to synchronize the heavy network traffic, the scenario based monitoring and partitioning of the network is required.

Scientific Experiments Designing Framework (SEDF)

Weather and climate researchers always need to design experiments for better understanding of the atmosphere. These experiments generally involve the statistical analysis of different observed parameters, data mining using combination of different parameters, etc. This framework under C-DME will facilitate the weather scientist to orchestrate the desired experiments. Once the experiments are set by the user, an automated workflow is generated and executed on the underlying software and hardware artifacts.

Work Flow Orchestration Framework (WOF)

This workflow designed by the weather scientist / operational forecasters using SEDF gets executed through this framework. This framework interacts with underlying heterogeneous hardware and software computational environment. The user can interact with this framework through a web portal. The workflow orchestration as per user requirements can be designed to overcome possible deadlocks due to global transaction failures or failure in the software environmental setting. As a corrective action for any of such failure detected in the workflow designing can either be communicated to the user or can be tracked by error tracking framework (ETF) for its automated debugging in real-time.

Scientific Data Manipulation Framework (SDMF)

Data manipulation is a prime concern of the weather scientists. This data manipulation framework under C-DME includes scientific data analysis, data processing for decision support, data visualization, data exploration, data archival and data interoperability components. These components use a unified approach to bring about their interaction with hardware/ software infrastructure as well as with other components of C-DME. In weather and climate modeling, the pattern of weather parameters differs and this framework can facilitate the user to employ statistical techniques such as data clustering, partitioning, pattern recognition and cross-correlated pattern detection. This framework is also helpful for data exploration, mathematical model building, its validation and verification, and finally its deployment to generate the predictions.

Scientific Data Visualization Framework (SDVF)

The availability of metadata of different forecast products, their tree based categorization and systematic archival of weather/ climatologicall data is useful for efficient and easy access to weather information. 3D geo-referenced data visualization (through the use of 3D graphics libraries such as VisAD) can support a researcher's natural strength of pattern recognition associated with the evolving weather. These 3D graphics as well as related data can be made available through a web-based interface or standalone API based interface to avoid transfer of large datasets over the internet. The different approaches using programming tools such as C, C++, VRML, Java, Java3D and ActiveX are required for platform

independence in the web-based visualization of 3D graphics. The use of client-server based architecture of the data server will help a forecaster to visualize weather data even from a remote location with proper authentication.

Content Management Framework (CMF)

The availability of metadata of different datasets is critical for efficient cataloging and indexing of the weather information and its efficient retrieval in an query based system. This framework would be associated with the query system interface for metadata as well as the actual weather/ climate data. The different functionalities of this component include data browsing, data transfer, data documentation, metadata organization at several levels, metadata creation, quality control, cataloging and indexing. The content management framework also supports the resource monitoring and metadata generation for the availability of hardware resources. Therefore, its design includes the low level interface that caters to the needs of data communication and storage. Further, as we are dealing with the huge datasets and large number of files, the high level I/O interface associated with the CMF perform all necessary I/O optimizations to improve the access to data through comprehensive discovery of requirement based metadata. The Lustre filesystem (Lustre, 2007) is designed for network-centric data storage, reliability and high-performance storage. This system properly authenticates clients and can provide a very good performance for I/O operations depending on the volume of data and time availability. The off-the-shelf component such as Scientific Data Manager (No et al., 2000) can be used to efficiently archive parallel/ Massage Passing Interface (MPI) I/O. C-DAC has developed the Grid File System (C-GFS) for sharing data storage capacity over the network. This system is designed for providing efficient and transparent access to massive storage space across heterogeneous environment and data distributed over wide area network no matter where it is physically stored in a grid. It also caters to data security and can be effectively utilized for data management across the grid environment.

Self Healing Framework (SHF)

The data volumes are approximately doubling each year and storage needs may reach to petabyte levels by 2010 AD. Such large scale of data volume requires thousands of hard drives and hundreds of nodes managing them, which can lead to inherent problems related to node failures, technological updates and incompatibilities between resources. To overcome this problem, the C-DME will have a reliable mechanism in place to protect against loss of data. This system shall work even though a part of it has failed and shall be associated with a self-healing system. Geographic replication of the data provides efficient data availability and protection against data loss. The design of the database management system shall comprise of an advanced data partitioning strategy to protect data against failures. The self-healing system will minimise data loss and facilitate dynamic correction to get the system back on track immediately during any such failure.

Real Time Load Balancing (RTLB)

Real time load balancing is one of the prime requirements of database query services to avoid delays in response. This demands proper resource scheduling,

partitioning and dynamic allocation of available/ additional resources especially to cater to the need during busy schedules. Crivelli and Head-Gordon (2004), have described a new load balancing strategy for improving efficiency of hierarchical approach of data partitioning for coarse grained problems associated with large tree searches. This strategy incurs minimal overhead and is scalable. The load balancing is achieved by reassignment of dynamical computational load tree based on the task granularity. Such strategies are directly applicable to the different scenarios that may arise in the scientific management of large database system as they are pattern based and use graph theory for data structuring.

Real Time Fault Tolerance (RTFT)

Real time fault tolerance at hardware level is essential to improve the availability of the system in case of any failure of hardware. Since a real time weather forecasting system implies the provision of reliable and timely information to weather forecasters, the consideration of fault tolerance is of utmost importance. Therefore, the design shall include a backup system to handle malfunctioning of any hardware device/ software code. The transit faults such as failure of network link between two sites can be overcome by providing atmospheric datasets at alternative locations. However, in case of the intermediate and permanent faults related to failure of one or more hardware systems, an automated rescheduling of processes among available resources is required. The use of geographically distributed resources is useful for achieving redundancy in the automated workflow execution and the selection/ rescheduling of the resources. The information redundancy can be achieved by replicating active memory multiple times to backup the active information. The temporal redundancy can be achieved by replicating similar operations of forecast prediction for several times. For example the use of ensemble technique for generating weather forecast may be useful to achieve temporal redundancy, since at least a few members among the ensemble will be continuously available to generate forecast.

Error Tracking Framework (ETF)

This framework generally tracks different user errors as well as exception failures generated by system. It communicates the errors to corresponding components for their automated debugging.

Data Mining and Data Warehousing Framework (DMF)

The data mining and data warehousing framework is designed to provide support to weather scientists and operational meteorologists for accurate result interpretation as well as for decision support. This framework generally supports creation of histograms, clustering of the data of different scenarios, various correlated pattern/ simultaneous patterns available in different weather parameters. This framework also supports the validation/ cross checking of the results obtained from scientific experiments. The availability of different techniques of predictive data mining such as feature selection, machine learning and meta-learning can facilitate the user to build actual model using workflow orchestration. The

structuring, organizing and partitioning of large data as an analytical process leads to development of flexible, efficient and open architecture data warehouse for weather and climate modeling.

Knowledge Discovery Framework (KDF)

This framework supports the development of generic tools to enable the users e.g. decision makers to understand the data. The development of a knowledge based expert system with available climatological data and weather / climate forecast climatology may help to enhance the utility of the system to cater to these requirements. Once developed, such system may be found to be useful for decision support related to estimation of forecast skills and value of forecast.

Grid Computing as an Alternative

The grid computing is a new paradigm for wide area distributed computing. This is a challenging area that promises the potential utilization of various resources available with institutions having diverse interests and spanning multiple administrative domains. This service provides flexible and dynamic access to computing resources through the web interface. The various components of this service enable access and discovery of the resources and datasets across heterogeneous networks. The Modeling Environment for Atmospheric Discovery (Wilhelmson et al., 2004; MEAD, 2007) expedition aims at the development and adaptation of Tera-grid enabled cyber-infrastructure for facilitating ensemble or very large domain model simulations coupled with data handling, analysis, data mining and visualization services. Linked Environments for Atmospheric Discovery (LEAD, 2007; Droegemeier et al., 2007) is aimed at the identification, providing access, data generation, data assimilation, weather predictions, weather data management, data analysis, data mining and visualization of a broad array of meteorological data and model outputs independent of format and physical location. These infrastructures provide an opportunity for the development of integrated scientific database system that is geographically distributed across the network. C-DAC has initiated grid computing in the form of the GARUDA (GARUDA, 2007) infrastructure. This initiative aims at to provide the technological advances required to enable data and compute intensive science for research scientists in multidisplinary areas. One of GARUDA's most important challenges is to strike the right balance between research and the daunting task of deploying resources into some of the most complex scientific and engineering endeavors. The C-DME can be considered as a one of the multi-institutional task under the aegis of GARUDA, India.

CONCLUSIONS

Weather and climate modeling requires the handling of large datasets which are acquired from different platforms using different observational techniques. These datasets may be in the form of model inputs, datasets for data assimilation, raw or processed outputs from global as well as regional models, visualization data files,

metadata, or data interpreted for its socio-economic benefits. The management of these datasets is an essential but daunting task. In this paper, we have presented a comprehensive scenario of the scientific data management for weather and climate research and proposed the interdisciplinary Data Management Environment C-DME to address the difficulties faced by the Atmospheric Sciences researchers. This system will facilitate the users with several functionalities not only for design of an experiment and its execution but will also take care of the underlying hardware and software heterogeneities, provide real-time self healing fault tolerance along with dynamical load balancing for efficient data management. The national initiatives such as GARUDA can provide a suitable environment to achieve the desired multi-institutional multi-disciplinary objectives for weather and climate related scientific data management.

Acknowledgments The help and suggestions provided by our colleagues at Computational Atmospheric Sciences, Scientific and Engineering Computing Group, CDAC, are thankfully acknowledged.

REFERENCES

Brown, A.W. and Wallnau, K.C. (1996) Engineering of component based system. In: Proc. 2nd IEEE International Conference on Engineering of Complex Computer Systems (ICECCS'96), 414-422. http://doi.ieeecomputersociety.org/10.1109/ICECCS.1996.558485 (accessed 2007).

Crivelli, S. and Head-Gordon, T. (2004) A new load balancing strategy for the solution of dynamical large-tree-search problems using a hierarchical approach. IBM J. Res & Dev., 48(2), 153-160.

Droegemeier, K., Baltzer, T., Wilson, A., Ramamurthy, M. and Lawrence, K. (2007) A New Paradigm for Mesoscale Meteorology: Grid and Web Service-Oriented Research and Education in LEAD. 23rd Conference on IIPS http://ams.confex.com/ams/87ANNUAL/techprogram/paper_117583.htm.

GARUDA (2007) GARUDA: India's National Grid Computing Initiative. http://www.garudaindia.in

HDF (2007) Hierarchical Data Format. http://hdf.ncsa.uiuc.edu.

LEAD (2007) Linked Environments for Atmospheric Discovery. http://lead.ou.edu/tech_links.htm. (accessed, 2007).

Lustre (2007) Lustre: A Scalable, High-Performance File System Cluster File Systems, Inc., White Paper. http://www.lustre.org/docs/whitepaper.pdf (accessed, 2007).

MEAD (2007) www-fp.mcs.anl.gov/pdq/MEAD.htm (accessed 2007).

M&D (2007) Model and Data Homepage, www.mad.zmaw.de (accessed 2007).

NETCDF (2007) Network Common Data Format, Unidata, UCAR, USA, http://www.unidata.ucar.edu/software/netcdf/

No, J., Thakur, R. and Choudhary, A. (2000) Integrated Parallel File I/O and Dataset Support for High-Performance Scientific Data Management. In: Proc. IEEE/ACM SC2000 Conference, 1-14.

Parnas, D.L (1972) On the criteria to be used in decomposing system into modules. Communications of ACM, 15(2), 1053 – 1058.

Ratnam, J.V., Sikka, D.R., Kaginalkar A., Kesarkar, A., Jyothi, N. and Banerjee, S. (2007) Experimental Seasonal Forecast of Monsoon 2005 using Global T170L42 on PARAM PADMA, accepted for publication PAGEOP.

Sha, L., Rajkumar, R. and Gagliardi, M. (1995) A software architecture for dependable and

evolvable industrial computing systems. Technical Report CMU/SEI-95-TR-005, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

W3C (2007) HTTP - Hypertext Transfer Protocol, http://www.w3.org/Protocols/ (access 2007).

Wilhelmson, R., Alameda, J., Droegemeier, K., Folk, M., Fowler, R., Gannon, D., Graves, S., Haidvogel, D., Husbands, P., Isbell, C.L. Jr., Weber, D., Woodward, P., York, B. W., Anderson, S. Jewett, B., Moore, C., Nolan, D., Porter, D., Semeraro, D. and Tanner S. (2004) MEAD (Modeling Environment for Atmospheric Discovery). In: Proc. 20th Int. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology (http://ams.confex.com/ams/84Annual/techprogram/paper_73057.htm).