Proc Indian Natn Sci Acad **80** No. 3 September 2014 pp. 697-704
© *Printed in India*.
DOI: 10.16943/ptinsa/2014/v80i3/55144

Review Article

Atmospheric Data Collection, Processing and Database Management in India Meteorological Department

A K JASWAL, N M NARKHEDE and RACHEL SHAJI
India Meteorological Department, Pune 411 005, India

(Received on 8 October 2013; Revised on 4 March 2014; Accepted on 6 April 2014)

The availability of a proper meteorological database is a major prerequisite for studying the processes of climate by scientists and policy makers. The acquisition of pertinent meteorological data, timely access and database management are important components that make the climate information valuable. New technologies in the field of communication and great strides in the automation age have offered new ways of making data products available to the user community. Data collection and database management are the key components of the availability of climate data. Due to increased observation frequency, huge amounts of data are received at data centres. Fast and effective quality control for identification and flagging of suspicious observations is needed at the data centres to provide easy access to information and dissemination of quality assured observations to the users.

India Meteorological Department (IMD) collects, processes, archives and disseminates a wide range of weather, climate and other environmental data generated by the department's observational network.With the installation of 'CliSys' Climate Database Management System (CDMS) in the National Data Centre located at Pune, the data acquisition and processing operations have been automated. This system provides a set of tools and procedures that allow all data relevant to climate studies to be properly stored and managed. It serves the main objectives of climate data collection, storage, quality control, easy access, data protection, climate data analysis and customized product generation.

Key Words: Database; Data collection; Data Management; Metadata; Quality Control

Introduction

India Meteorological Department (IMD) is the National Meteorological and Hydrological Service (NMHS) of India responsible to take meteorological observations and to provide information for weather sensitive activities. All national climate activities, including research and applications are primarily based on observations, recording the state of the atmosphere. These observations are also required for the timely preparation of weather and climate analyses, forecasts, warnings, research and other national and international environmental programmes. The importance of meteorological data for operational needs of weather forecasters, researchers and crop forecasters are gaining attentions in the domestic and the international research community. Data acquisition, processing, quality control, timely access and database management are important components that make the meteorological information valuable in issuing weather forecasts as well as climate and agricultural research. The increasing challenge is to manage these growing resources of data so that users can get the needed information for their specific applications. Computer technologies are essential tools for processing and archiving huge amount of meteorological data and providing derived information to all types of users. In order to meet the new challenges in managing atmospheric data, climatological data processing

**Author for Correspondence: jaswal4@gmail.com*

activities at the National Data Centre (NDC) in IMD were upgraded with a new Climatological Database Management System (CDMS) named 'CliSys' which has been developed according to the World Meteorological Organisation (WMO) guidelines by Meteo France International (MFI). This is a new step to secure the climate data heritage of the country and manage the atmospheric data efficiently. CliSys is a set of tools and procedures that allow all data relevant to climate studies to be properly stored and managed. This paper outlines the current methods of atmospheric data collection, processing, quality control and database management in IMD making use of the newly acquired CDMS.

Methods of Data Collection

Surface weather observations are fundamental to all meteorological services upon which forecasts and warnings are made in support of wide range of weather sensitive activities. Collection of meteorological data electronically atthe source allows automatic quality controls to be applied, including error checking, prior to the data transmission from the observation site. The methods of observations in IMD can be categorized into two major classes, manually observed and automatic data collection stations. Manual observations are recorded in the manuscript form at the observing stations which are sent to the designated Meteorological Centers (MCs) or Regional Meteorological Centers (RMCs) as shown in Fig. 1. After manual scrutiny of these data at RMCs/MCs, they are keyed-in in the well defined formats and sent to National Data Centre (NDC). Data generated by automatic weather stations are subjected to minimum quality checks before transmitting through Automatic Message Switching System (AMSS) and are archived at NDC in near real-time basis. The processes involved in data acquisition, processing, quality control and dissemination in IMD are shown in Fig. 2.

Manually Observed Data

Manual surface observatories are located almost one in each district so as to meet the requirements of weather forecasting, agriculture and other operations of the country. There are at present 530 surface observatories, 45 radiation observatories, 10 air pollution monitoring observatories, 62 Pilot Balloon observatories, 39 Radiosonde and Radiowind observatories, 219 agro-meteorological observatories and more than 700 hydro-meteorological observatories. In addition to this, marine surface data are also recorded and transmitted to IMD by more than 200 ships of the Indian Voluntary Observing Fleet (IVOF). Of the 530 surface observatories, nearly two third are manned by the non-departmental staff. The observers of non-departmental observatories are given training in taking observations while the necessary instruments and other stores are provided by the IMD. The meteorological data thus collected from all over the country are used on real time basis for operational weather forecasting. These data records arefurther quality checked and archived in NDC for various kinds of usage including climate related studies.

Manually observed data forms the bulk of meteorological database in IMD which is used in preparing meteorological information not only in the planning but also in the operational fields. The observations taken manually are air temperature, relative humidity, sunshine duration, solar radiation, wind speed and direction, precipitation, evaporation and cloud cover, etc. The main mode of transmitting data from observation sites to the RMCs and MCs is by internet and telephones. The surface and upper-air observations in World Meteorological Organization (WMO) standard coded messages are sent through Global Telecommunication System (GTS) for operational use and real-time data processing, quality control and archival use. However, all meteorological observations recorded by observatories in IMD's network do not flow through GTS. Instead, these observations are manually scrutinised and keyed in at RMCs and MCs in delayed mode in standard formats supplied by NDC. These data files are then transferred to NDC for further processing and archival use. The availability of various kinds of data archived in NDC is given in Table 1.

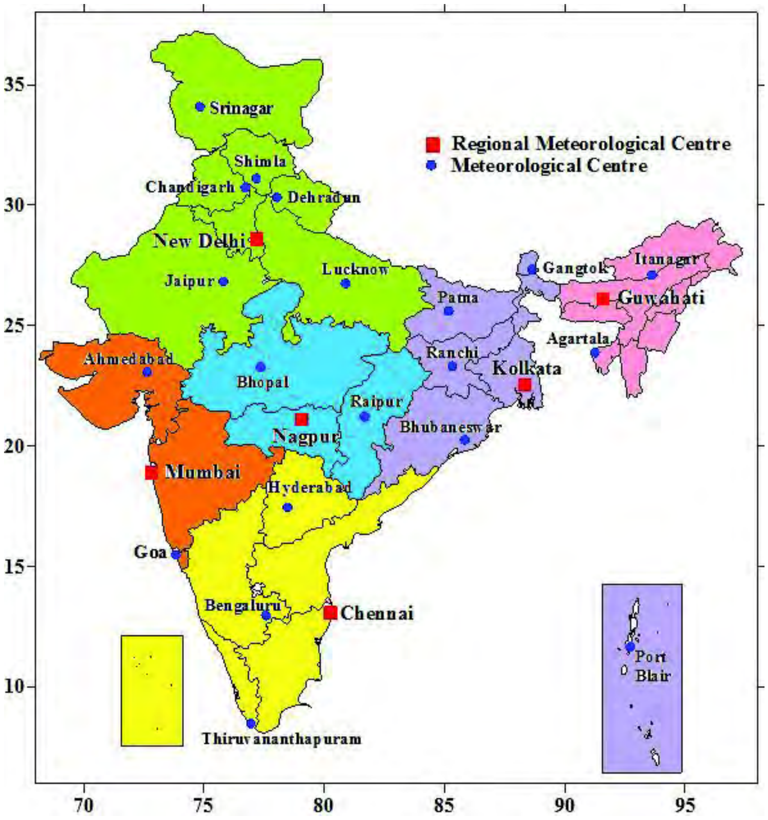

Fig. 1: Location of Regional Meteorological Centers and Meteorological Centers

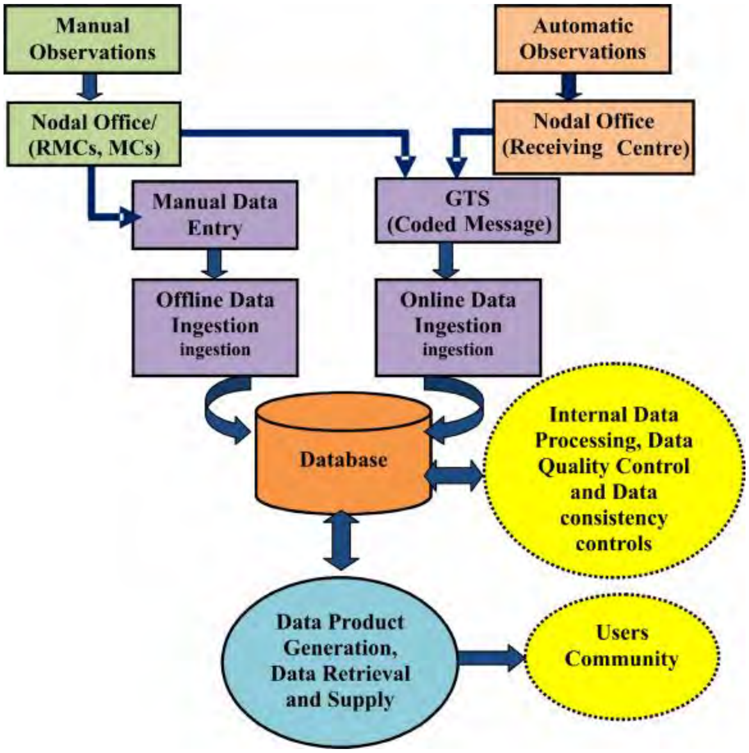
Manual Observations
Automatic Observations
Nodal Office/ (RMCs, MCs)
Nodal Office (Receiving Centre)
Manual Data Entry
GTS (Coded Message)
Offline Data Ingestion ingestion
Online Data Ingestion ingestion
Database
Internal Data Processing, Data Quality Control and Data consistency controls
Data Product Generation, Data Retrieval and Supply
Users Community

Fig. 2: Schematic view of data reception, processing and quality control

Automatic Weather Data

Automatic Weather Station (AWS) and Automatic Rain Gauge (ARG) data are increasingly becoming the ideal method for collection of meteorological data world over. The numbers of AWS and ARG stations are growing at a rapid pace in IMD also. At present, more than 650 AWS and 1450 ARG stations have been installed all over the country and these stations are working under IMD's network. The AWS and ARG data are collected at hourly frequencies and transmitted to the data receiving center located at Pune.These data are automatically quality checked before transmission on GTS for operational use.The parameters collected are air temperature, relative humidity, wind speed, barometric pressure, rainfall, soil temperature, leaf wetness and soil moisture. The CDMS in NDC imports AWS/ARG data through AMSS for online data ingestion into the database. These data are further put through in-built quality control procedures of CDMS.

Climate Data Management System in IMD

In order to manage the atmospheric data efficiently, IMD has created a Climatological Database Management System at National Data Center, Pune. The management of climatic data in this IT age has become easier, faster, and more efficient. The key areas in managing climate information include data collection, its processing, quality control, analysis, product generation and product delivery. 'CliSys' is a Climate Database Management System installed in National Data Centre, Pune where all atmospheric data are archived. It is designed to ingest online data through

Table 1: Atmospheric data availability in the archives of National Data Center (NDC). The availability of a particular parameter depends upon its year of start of recording.

Data type	Data frequency	Availability (year onwards)	Parameters
AWS-SYNOP	Hourly	2007	Atmospheric pressure, Air Temperature, Dew point temperature, Rainfall, Wind (Direction and Speed), Sunshine hours
SURFACE	Daily	1969	Atmospheric pressure, Air temperature, Dew point temperature, Humidity, Vapour Pressure, Evaporation, Rainfall, Wind (Direction and Speed), Sunshine hours, Visibility, Weather phenomenon
SURFACE	Monthly	1901	Atmospheric pressure, Air temperature, Dew point temperature, Humidity, Vapour Pressure, Evaporation, Rainfall, Wind (Direction and Speed), Sunshine hours, Visibility, Weather phenomenon
RAINFALL	Daily, Weekly, Monthly	1875	Rainfall
UPPER AIR	Daily, Monthly	1951	Air Temperature, Dew Point Temperature, Wind (Direction and speed)
AUTOGRAPHIC	Hourly	1969	Atmospheric pressure, Air temperature, Humidity, Wind speed, Pressure, Rainfall, Sunshine
MARINE	Hourly	1961	Atmospheric pressure, Air temperature, Dew point temperature, Sea surface Temperature, Present and past weather, Wind (Direction and Speed), Visibility, Clouds, Wind wave, Swell wave
AGROMET	Hourly	1972	Air Temperature, Wet bulb temperature, Relative humidity, Vapour Pressure, Evaporation, Evaporation transpiration, Rainfall, Wind (Direction and Speed), Sunshine hours, Atmospheric pressure, Soil temperature, Soil moisture
RADIATION	Hourly	1957	Global, Diffused, Direct, Net, Terrestrial
TURBIDITY	Daily	1980	Turbidity
OLR	Daily	1987	Outgoing Longwave Radiation
OZONE	Weekly	1980	Ozone
MISCELLANEOUS			MONEX-79, MONSOON-77, ISMEX-73, Special Expeditions, etc.

AMSS and offline loading by assigning semi-colon delimited ASCII files. The system includes data entry, data monitoring, data retrieval, reporting, quality control and metadata sub-systems integrated into a comprehensive CDMS.The system has been designed to operate in a three-tier configuration viz. Database server, Application server and Web server. The system is mainly composed of five subsystems namely the Database (in-charge of the storage of climate data), the Data Acquisition (responsible for real time and non-real time acquisition of climate data), the Metadata Management (takes care of the acquisition and management of metadata), Data Management (responsible for the climate data management) and Production (in-charge of data products elaboration and data access) as shown in Fig. 3. The main features of CDMS installed in NDC are:

A unified Data Storage Structure

It is built around a Relational Database Management System (RDBMS) where data, including all the relevant metadata, are stored in one unique powerful database ensuring thereby the centralization and uniqueness of all the information. Furthermore, the features offered by RDBMS ensure, among others, performances in climate data retrieval, reliability through various back up mechanisms, climate data

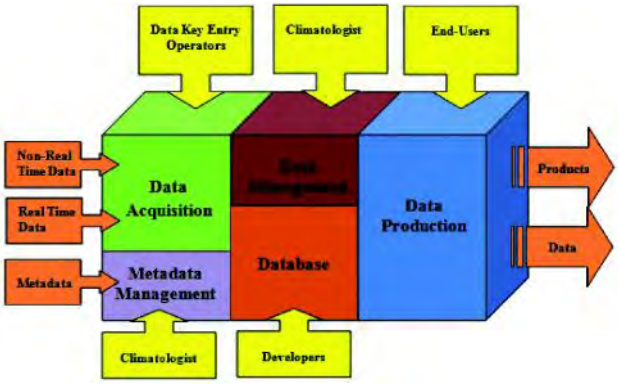

Fig. 3: Schematic functional view of Climate Data Management System

consistency, integrity constraints and concurrency controls as shown in Fig. 4.

A Dynamic Data Structure

It is based on a data model defining the logical storage structure specifically designed for climate data in order to guarantee high performance for real time applications and to optimize data storage.This model is built dynamically in order to meet the requirements to store any kind of meteorological information.

A Flexible and Open Data Import System

'CliSys' is able to ingest all kinds of climate data as it has flexible and open data loading system. It provides methods to acquire climate data ranging from historical data already in place, to real-time acquisition from AWS/ARG or GTS via manual key-in entry processes.

A Web Based Architecture Allowing Regionalisation/Decentralization

The CDMS is based on a web based technology. Through a simple internet browser, user-friendly interfaces and user rights policy, it offers an easy and instantaneous access to the system administration, the data management or the products generation.

Task Management in Clisys

On the basis of different tasks to be performed, the CliSys CDMS can be viewed as shown in Fig. 5. The classification of various tasks is done by the nature of task to be performed. These tasks are broadly classified as Metadata Administration, Computation, Quality Control and Operations. While metadata administration consists of activity of gathering correct and complete information about observatories, data collection, and chronological information related to

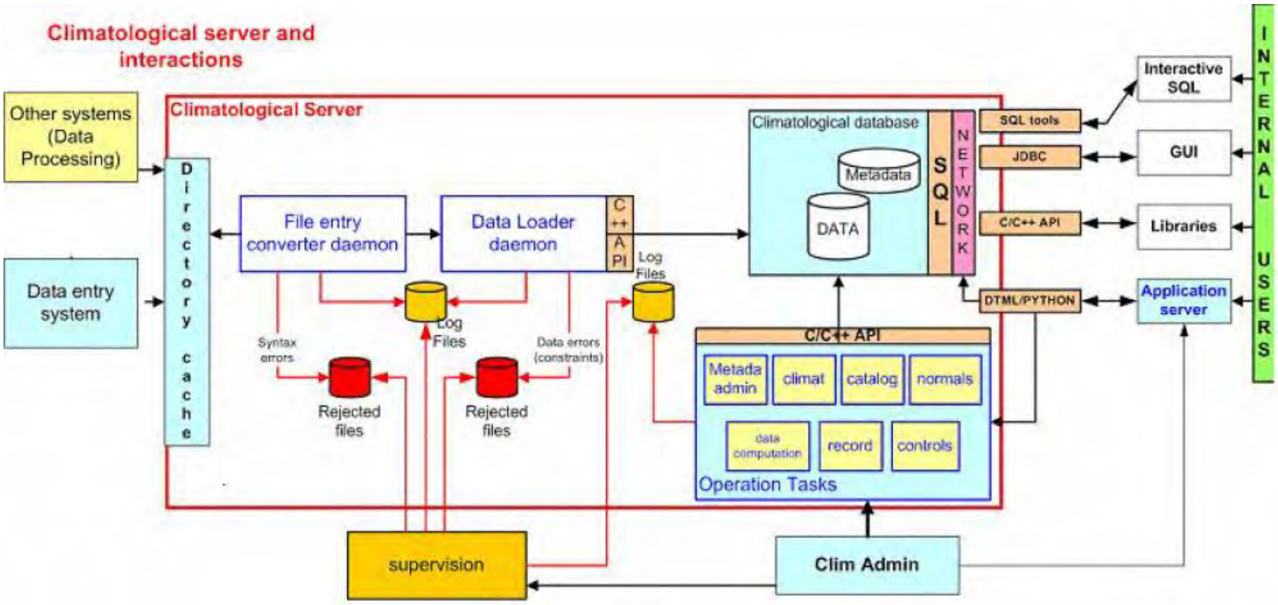

Fig. 4: Data consistency and integrity controls in Climate Data Management System

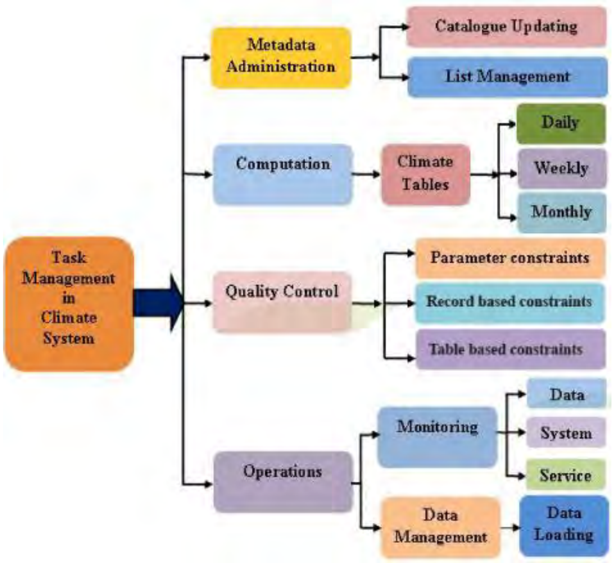

Fig. 5: Schematic functional view of Task Management in Climate Data Management System

changes in the observatories, data formats for storing data and its domains (data types), etc., it ensures the referential integrity of data and avoids redundancy and ambiguity of data. Tables storing this vital information about the data (metadata) are referred to as catalogue (Directory of directory). The information can be stored as requirements in groups as listing (state wise, subdivision wise lists, etc.). Computation is an important activity responsible for automated operations of system components to achieve desired goals. It focuses mainly towards extraction of data from network (GTS, AMSS, etc.) and puts it in structured tables of database with recommended norms. This also includes computation of many derived parameters and tables on the fly. Quality control includes applying standard algorithms to check the authenticity of extracted data value. This also contains database related constraints to ensure uniqueness of data records and various relational aspects of parameters within records. Operations consist of many activities to run and get maximum throughput of the system. It includes monitoring the system for availability of services and data 24x7 and data loading of historical data in the system with matching data loading formats.

The data catalogue is generated automatically giving information about the present and missing data from the data tables for each of the parameters of these tables. It helps in preparing data estimates in answering queries from various users.

Data Reception and Processing

Real-time Data

Real-time data are observations retrieved by automatic data processing methods from the site in real or almost real time and transmitted to a data collection centre instantaneously and delivered to real-time data users, e.g. weather forecasters. Real-time automatic acquisition processes in CliSys consist of two main background modules. The first module reads and converts input ASCII files in CliSys format already described. The second one inserts these files into the database. Real-time meteorological observational data such as synoptic messages and AWS/ARG data flowing through GTS are decoded to the proper variables of the CliSys database and are archived. These data consist of WMO standard coded messages viz. SYNOP, AWS, ARG, TEMP, METAR and SHIPS. Data values are checked for its global limits and internal consistency before ingestion into the database.

The real-time data reception in the database is visualized in near real time basis. A geographical data monitoring displays the IMD's observation network on map of India and related status of the incoming data from stations through colour codes. Also a tabular form indicates the status (missing value rate) of a given parameter for a given period and for a given station, in accordance with the theoretical metadata frequency measurements. It displays the real/expected time series in both tabular and graphical forms.

Non-Real Time Data

The non-real time data are those from the manual stations which are sent in a delayed mode and received from the Regional Meteorological Centres (RMCs) and Meteorological Centres (MCs) of IMD. These consist of surface, autographic, upper-air, radiation, rainfall, agro-meteorological data. All observatories send their observation records to their respective data-keying centres on monthly basis. These observations are first manually scrutinized by the technical sections of the respective data-keying centers located at RMCs and MCs and then keyed-in in the pre-defined NDC supplied data formats. After receiving these keyed-in observations from the data-keying centres, quality checks viz. valid character check for each field, duplicate checks, extreme value limit, internal consistency, hydrostatic checks for the upper-air data are applied at NDC as per the guidelines given by the WMO for data processing (WMO 1992). These data in ASCII format are then converted into the CliSys format and are ingested into the database. The non-real time data acquisition processes use the same data loading procedures as used by the real-time data.

Quality Control

An important part of the management of an NMHS is the data quality control process in operation. Due to station automation, increased observation frequency and increased data transmission speeds, huge amounts of data are received in data centres world over. The quality control process is designed to check the quality of the whole data-flow ensuring that data is error-free as far as possible. Therefore, fast and effective quality control for identification and flagging of suspicious observations is needed to provide easy access to correct information and dissemination of reliable observations to the users. There are several points where errors can creep into the data and so these must be detected and eliminated, and if possible, the errors should be replaced by the corrected values while also retaining the original values. World Meteorological Organisation (WMO 1981) prescribes that certain quality control procedures must be applied to all meteorological data for their international exchange.

Quality Control Requirements

The main purpose of quality control is to check whether a reported data value is representative of a real phenomenon, or is an outlier not consistent with present weather conditions. Data errors can enter the data set at many sources including the observation site, instrument/sensor, data transmission or data entry stages. All data quality checks and flagging system in the CDMS have been developed as per the WMO guidelines (WMO 1993).Quality controls for each data are mainly designed for hourly and daily meteorological data in the system. They make real-time checks at the data ingestion level through tolerance tests. After data loading, the data control process is run with results stored in a doubt table and the quality control flags are updated. The quality controls pass through four levels which are Syntax cum Tolerance tests, Priority test, Filter tests (Oracle table constraints) and consistency controls. The “Meteorological” controls are performed once the data are in the database. These are Global limit check, Related elements global limits, Internal Consistency, Global rate of change and station record limits, Standard deviation check and Station rate of change check. Further, spatial consistency checks can be applied by visualization of climate values spatially with direct link to data modification functionalities. Once the data are in the database, non-real time tests consisting of internal consistency i.e. physical relationships among climatological elements and temporal consistency i.e. variation of a climatological element in time are applied.

Quality Control Flags

From the data ingestion to the archiving processes and subsequently through the data management module, the CDMS checks the validity of a data and flags it with the appropriate coded value. During quality control, suspicious values or certain errors are identified and quality control flags are associated to each climate element. Flagging information is assigned to each data element in order to indicate the level of data quality. However, some erroneous and suspicious values are inspected manually to avoid exclusion of extreme weather phenomena.

Metadata Management

Metadata is the information about the data that helps in its understanding and use. It is the descriptive information necessary to allow users to find, process and use the data, information and products. Metadata are various kinds of information about meteorological stations such as general station information (e.g.

geographical co-ordinates, address, etc.), station history, sensor history (e.g. calibration, repairs, etc.), station environment (e.g. terrain, exposure, etc.), kind of equipment, state of station, sensor statistics (e.g. frequency and kind of error, etc.), service information and plans (WMO 2003). CliSys provides a module to specifically manage the metadata. It has features of adding new values (stations, instruments, etc.) and modification of existing values. It has in-built tools to consult and display the metadata values. It uses more than 50 tables to manage stations details, instruments, parameters, geographic characteristics, sensors information, site pictures, etc. Metadata have a key role in the process of creating datasets, as the knowledge of the station history provides increased confidence in the statistical techniques employed to ensure that the only variations that remain in a climate time series are due to actual climate variability and change. Most metadata have been derived from the station's documentation, both from current and historical documents available in the department.

Data Retrieval and Supply

With consideration of data policies existing in IMD, a variety of climatological data both basic and derived products are regularly supplied by the NDC in response to a large number of queries received from central and state governments, universities, research institutes, public sector undertakings and private sectors. These data queries are categorised as departmental, research institutes and commercial organisations. Research institutes can register with IMD giving specific details of the projects to get meteorological data at very nominal rates. The data supply to commercial organisation is done by charging actual cost of data which is revised after every three years. The information supplied is used for lay-out of run-ways, town planning, air-conditioning, industry, port installations, installation of high towers, bridges and other structures, operation of multipurpose energy projects, water and power management, defence operations in inaccessible regions, calibration of defence equipment, environmental studies, renewable energy sources and many more. Daily AWS/ARG data are freely available for general public on IMD Pune website (http://imdpune.gov.in/aws/aws_index.html). Monthly extremes of temperature and rainfall for surface observatory stations under IMD network are compiled every month and are also hosted on IMD Pune website for the benefit of all users.

Summary

Understanding and predicting changes in climate are important activities for any NHMS of a country. With the development of information technologies, the research on weather and climate has also made a great progress. The power and user-friendliness of computers in collection, transmission, processing and storage of meteorological data and the ability to record and transfer information electronically have given climatologists new tools to rapidly improve the understanding of climate. Hence, it isessential to archive and provide access to the reliable climatological data needed to describe, understand, and predict changes in the Earth System as well as the impacts of these changes on society. With WMO Information System (WIS) architecture at national level, better data management and storage will allow easy discovery, better access and retrieval of climate information and services in future.

Acknowledgements

Authors are grateful to Dr. L. S. Rathore, Director General of Meteorology for encouragement and support. We are also thankful to the unknown reviewers for constructive suggestions that helped to improve this manuscript.

References

WMO (1983) Guide to Climatological Practices, Second Edition.WMO No.100, Geneva

WMO (1992) Manual on the Global Data Processing System, WMO No. 485, i Geneva

WMO (1993) Guide on the Global Data Processing System, WMO No. 305, Geneva

WMO (2003) Guidelines on Climate Metadata and Homogenization,World Climate Data and Monitoring Programme (WCDMP-53), Geneva.