

Development of Archival Precipitation Data Sets for the GCIP Domain

B. R. Nelson, W. F. Krajewski, A. Kruger
The University of Iowa, Iowa City, IA, USA

J. A. Smith, M. L. Baeck
Princeton University, Princeton, NJ, USA

ABSTRACT: We present the development of an archival precipitation data set for the GEWEX Continental Scale International Project (GCIP) region in the Mississippi River basin of the central United States. We propose to produce hourly rainfall maps on approximately a $4 \times 4 \text{ km}^2$ grid for the GCIP domain that will be used by the modeling community for a variety of projects. Furthermore, we propose to develop these hourly rainfall maps for the period of record of 1995 through 2000. The scope and magnitude of this undertaking is without precedent in the fields of hydrology, hydrometeorology, and meteorology. Earlier precipitation data sets, such as GATE, GPCP, or TOGA-COARE are much smaller. This data set will be developed from multi-sensor precipitation estimates. We propose to use the National Reflectivity Composite and hourly rain gauge data as the principal data sets. However, the NCEP hourly rainfall composite, the WSR-88D reflectivity data, and data from the GOES-8 imagery will supplement the principal data sets. As with any multi-sensor estimate of precipitation from radar and rain gauge data, issues such as quality control of the radar and rain gauge data and inter-comparison studies of radar-radar calibration differences, range dependence of radar rainfall estimates, and inherent biases in radar rainfall estimates will be investigated. This project demands additional investigations with respect to data set development and organization. We propose to determine a suitable compression and storage scheme specific to this data set and address the issues of visualization, portability, and distribution. We present the preliminary findings involved with the development of such a large data set: in particular, the investigation into the best compression and storage scheme for this data set.

1 INTRODUCTION

The concept of the Global Energy and Water Cycle Experiment (GEWEX) Continental Scale International Project (GCIP) was conceived in 1990 as the United States' contribution to the overall scientific strategy of the World Climate Research Programme (WCRP) and GEWEX. As the first of the five Continental Scale Experiments (CSE), GCIP was established to quantitatively determine the hydrologic cycle and energy fluxes of the

Mississippi River basin. The overall goal of the GCIP is to demonstrate skill in predicting changes in water resources on time scales up to seasonal and annual, as an integral part of the climate system. The development and evaluation of coupled hydrologic-atmospheric models at resolutions appropriate to large-scale continental basins are critical to the successful achievement of the GCIP goal (Coughlan and Avissar, 1996). Similarly, the development and evaluation of hydrologic and atmospheric

models at less than continental-scales are critical to the GCIP goal.

We propose to develop an archival precipitation data set for the GCIP region. We propose to produce hourly rainfall maps on approximately a 4×4 km² grid for the GCIP domain that will be used by the modeling community to achieve the GCIP goal. Hydrologists, hydrometeorologists, meteorologists, and atmospheric scientists will use this data set for meso and synoptic scale models. The key to the successful development of this large archival data set is the implementation of sound data infrastructures.

Issues arise when developing a data set at such a fine temporal and spatial resolution over a large basin for a 5-year period. One of the first issues is data compression. Furthermore, accessing the data is a major hurdle, and storing the data is an issue. Problems related to quality control of the data and validation and calibration will also be addressed in this paper.

2 DATA SOURCES

The data used as input for development of the 4×4 km² hourly rainfall maps will come from different sources. We will obtain data from the National Reflectivity Composite and hourly rain gauge data. These data will be our principle inputs for developing the rainfall maps. However, the NCEP hourly rainfall composite, the WSR-88D reflectivity data, and data from the GOES-8 imagery will supplement the principal data sets. The supplementary data sets will be used for quality control and validation and calibration of the compiled 4×4 km² hourly rainfall maps.

2.1 Radar Reflectivity Composite

The National Reflectivity Composite is derived from the National Weather Service (NWS) radars and is compiled by the Marshall Space Flight Center (MSFC) Distributed Active Archive Center (DAAC). The MSFC DAAC receives quality-controlled images and metadata of radar reflectivity from Weather Services International (WSI) Corporation. The

WSI Corporation creates merged radar reflectivity data from the NWS 10 cm radars that are operational in the continental United States (CONUS).

The MSFC DAAC receives instantaneous snapshot images at 15-minute intervals that are derived from the combined NWS radar data. The MSFC DAAC converts these images into a composite rainfall map for the CONUS. The 15-minute composite rainfall maps are stored in Hierarchical Data Format (HDF) as raster images for data compression. Two rainfall rate products are generated at MSFC DAAC. One product is the 15-minute instantaneous rainfall rates, and the other is the daily rainfall rate. The daily rainfall rate map is produced only if all 96 15-minute instantaneous files are present.

2.2 Rain Gauge Data

We will obtain rain gauge data from rain gauge networks in the CONUS. A variety of rain gauge networks will be identified for inclusion into the development of the rainfall maps. We will use networks such as the Oklahoma Mesonet and the North Dakota Rain Gauge Network operated by the North Dakota Resource Board, for validation studies of the radar reflectivity composites. We will compile other rain gauge data from the National Climatic Data Center (NCDC) for use in the validation phase of the project.

2.3 Supplemental Data

We will use supplemental data for additional validation of the rainfall maps. The NCEP hourly rainfall composites and level II volume scan data from specific WSR-88D radars in the CONUS will serve as additional data for validation. We will use volume scan data from Memphis, TN, Denver, CO, Pittsburgh, PA, Greer, SC, Tulsa, OK, Davenport, IA, Houston, TX, and Grand Forks, ND. Also the GOES 8-9 satellite data will serve as additional space-born data for validation.

3 DATA FORMATS

The selection of the format of the rainfall maps is more complex than it first seems. We need to consider several issues when deciding on a format for the final product of the rainfall maps. There are several different formats available for the storage of scientific data sets, and each of these formats have their advantages and disadvantages. We will present a few of these formats and some advantages and disadvantages with respect to our data.

3.1 Hierarchical Data Format (HDF)

The HDF format is a multi-object file format for sharing scientific data in a distributed environment. This format was created at the National Center for Supercomputing Applications (NCSA) to serve the needs of scientists working on projects in many fields. The HDF format was created to address many requirements for storing scientific data.

Among the many advantages of HDF are efficient storage of and access to large data sets, platform independence, and extensibility for future enhancements and compatibility with other standard formats. Disadvantages of HDF include the learning curve necessary to become efficient in using the libraries of functions and the computational time necessary to access an HDF file and read its encoded data. Furthermore, we question if it is necessary to obtain an entire set of libraries and binaries just to be able to access one type of file?

3.2 Common Data Format (CDF/NetCDF)

The CDF is a scientific data management package that allows programmers and application developers to manage and manipulate scalar, vector, and multidimensional data arrays (Mathews 1996). The Network Common Data Form (NetCDF) is an extension of the CDF that includes a machine-independent external data representation, a C interface, a Unix implementation, and a mechanism to access data aggregates in addition to single data values (Rew and Davis 1990).

Similar to HDF, the advantages of CDF and NetCDF are an application-independent interface to data, access to multidimensional data, and the representation of relations among the elements of multidimensional arrays. The disadvantages are again similar to HDF. The learning curve is steep to become an efficient user, and again, it is necessary to obtain a specific set of libraries and binaries to be able to produce and access one type of file.

3.3 ASCII-Run Length Encoded (ASCII-RLE)

The ASCII-RLE format allows for storage of scientific data in scalar, vector, or multidimensional arrays. This format uses a run-length encoding scheme for data compression and stores the data in an ASCII encoded format for easy manipulation (Kruger and Krajewski 1997). This format has been used for storage of the RADAP II data set and the TOGA-COARE data set (Kruger et. al 1999).

The advantages of the ASCII-RLE format are the highly compressed format significantly reduces the size of large scientific data sets, and the ASCII encoding of files makes for easy manipulation of the data. The disadvantages of this format are the data-set specific format and the relatively small user implementation as compared to HDF or NetCDF.

4 DATA STORAGE

Issues related to data storage are closely related to the selection of the data format. The size of the data and any data compression techniques will influence the selection of the storage media selected. We will investigate the coupling of these issues to determine a suitable format and storage scheme.

4.1 Data Size

The size of the rainfall data set in the raw format makes it necessary to investigate methods and file formats that will significantly decrease the total size of the data set. For instance the Mississippi River basin is approximately 3.6×10^6 km². At a resolution

of $4 \times 4 \text{ km}^2$ and for a 16-bit integer representation of the data, one rainfall map could be as large as approximately 1 Mb of data. If the data were represented as binary, one rainfall map could be as large as 0.25 Mb. Therefore, data for one year at a resolution of hourly intervals could be as large as 8 Gb in raw binary format. We have proposed to develop rainfall maps for the period of 1995-2000. Thus the size of the raw binary data set could be as large as approximately 40 Gb. Obviously, 40 Gb is much too large for the format of the rainfall maps to be distributed to the users. Therefore, we will investigate data compression techniques that will help to reduce the size of this data set.

4.2 Data Compression

The reflectivity composites distributed by the MSFC DAAC are stored using the HDF file format. The data are stored in a raster image format inside the HDF structure. Within this structure, the data are stored using a run-length encoding scheme. This format greatly reduces the size of the rainfall maps. For example, the size of the data set for one year is approximately 1.5 Gb. However, for the period of record, the size of the data set would still be approximately 8 Gb. Can we do better?

We have investigated converting the rainfall maps that are stored in the HDF format to the ASCII-RLE format. We have converted the HDF formatted data for one year to the ASCII-RLE format, and we have found a compression ratio of approximately 4 using the ASCII-RLE format. This format greatly reduces the size of the data set. The size of the entire data set for the 5-year period could be reduced to approximately 3-4 Gb using the ASCII-RLE format.

Figure 1 shows a histogram of the size of the reflectivity composite maps for the HDF file format. The files are 15-minute $2 \times 2 \text{ km}^2$ reflectivity composites for the warm season (March-October) of 1999. Figure 2 shows the histogram for the same files in ASCII-RLE format. The ASCII-RLE formatted files save a considerable amount of disk space, and as can be seen in Figure 3, the computational time needed to access an ASCII-RLE formatted file

is considerably less. The savings in computational time is shown for just one year of warm season data. This savings will increase for the five-year data set.

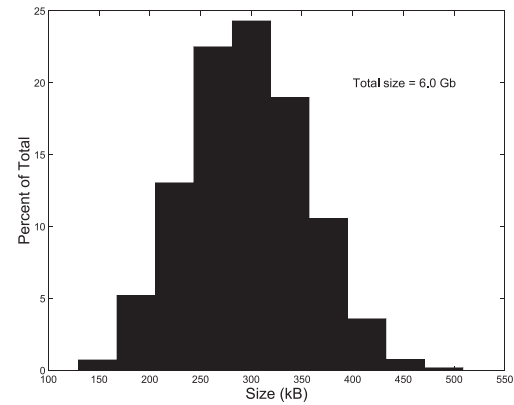


Figure 1: Histogram of file size for HDF file format of 15-minute $2 \times 2 \text{ km}^2$ reflectivity composites.

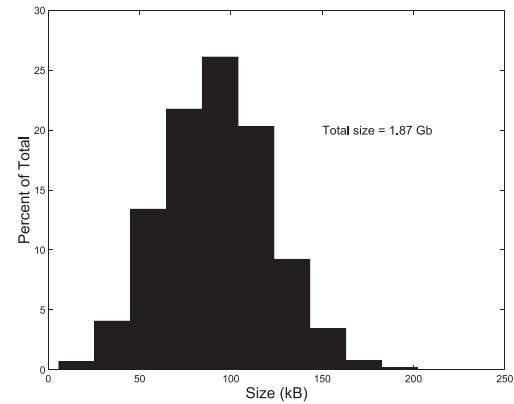


Figure 2: Histogram of file size for ASCII-RLE file format of 15-minute $2 \times 2 \text{ km}^2$ reflectivity composites.

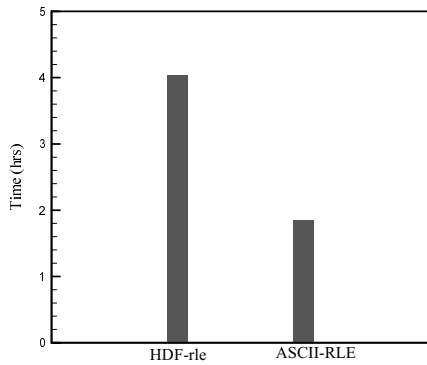


Figure 3: Computational time to access HDF and ASCII-RLE formatted files.

4.3 Storage Media

Another interesting problem we have found that is more complicated than first seemed is the selection of the storage media for distribution to the user. There are many ways for distribution of data but the size of this data set limits the possibilities. Devices such as CD-ROMs and DVD-ROMs are possibilities. Internet and ftp access are also possibilities. The limitation of CD-ROMs is the storage size. At 3-4 Gb needed for storage of the data set, we would need 5-6 CD-ROMs for distribution. The limitation behind DVD-ROMs is the development of the technology. It is still unclear as to whether or not the technology is a viable alternative for data storage and transfer. The limitation of internet or ftp access is the time to download or transfer data.

Initially, we believe the technology of DVD is such that the rainfall maps data set can be distributed on 1-2 DVD-ROMs. At this point, a 2-sided DVD-ROM can store 9 Gb of data. Therefore we anticipate that the data set could easily be transferred to the DVD-ROM and distributed as a 1-2 single or double-sided disc set.

5 PORTABILITY

We plan to produce the rainfall maps to be portable across platforms. The platforms that we are considering for portability include the

PC-Windows and the PC-LINUX environment, as well as the HP-workstation and Solaris-SGI, Unix workstation environments.

6 VISUALIZATION

Visualization is a key part of developing the rainfall maps for such a large area. Users of this data set will inevitably want to be able to view portions of the data set before implementing it into models. As the visualization is a key element in the development, it is also a difficult part of the development. We will investigate packaged software such as the Java HDF Viewer available from NCSA and commercial software such as IDL. We will also examine the technologies of geographic information systems (GIS). Finally, we will determine if we need to develop a data-set specific visualization software.

7 QUALITY CONTROL/VALIDATION

Initial investigations into the National Reflectivity Composite data show there are some data quality issues apparent in the composite reflectivity maps. We have found instances of ground clutter, anomalous propagation (AP), and beam blockage. Therefore, we will need to implement systematic techniques that can help to identify these problems in the input data. Systematic implementation of algorithms for identification of ground clutter and AP has been made by Moszkowicz et. al, 1994, Grecu and Krajewski, 1999, and Grecu and Krajewski, 2000. We will determine the applicability of these and other techniques for our implementation.

Additionally, calibration and validation of the hourly 4×4 km² rainfall maps will be important. We will use rain gauge networks in the CONUS and satellite data for validation and calibration. We will use specialized local networks such as the Oklahoma Mesonet and the Goodwin Creek Research Watershed (Steiner et. al 1999).

8 CONCLUSIONS

We have proposed to develop rainfall maps for the GCIP domain in the Mississippi River basin of the CONUS. The maps will be developed at a spatial resolution of 4x4 km² and a temporal resolution of one-hours. The period of record for the rainfall maps is from 1995-2000.

The rainfall maps that we will produce will be used in a variety of applications and by a variety of users. We will produce a quality product that can be implemented in many fields. This product will be used for hydrologic modeling, climate modeling, and water and energy budgets to name a few. Therefore, we need to address the issues of data set size, data quality control and data validation.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of NOAA Office of Global Programs through grand NA96GP0417.

REFERENCES

- Coughlan, M. and R. Avissar, "The Global Energy and Water Cycle Experiment (GEWEX) Continental-Scale International Project (GCIP): An Overview," *Journal of Geophysical Research*, **101**(D3), pp. 7139–7147, March 20, 1996.
- Greco, M. and W.F. Krajewski, "Detection of Anomalous Propagation Echoes in Weather Radar Data Using Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, **37**(1), pp. 287-296, January 1999.
- Greco, M. and W.F. Krajewski, "An Efficient Methodology for Detection of Anomalous Propagation Echoes in Radar Reflectivity Data Using Neural Networks," *Journal of Atmospheric and Oceanic Technology*, **17**, pp. 121-129, February 2000.
- Kruger, A. and W.F. Krajewski, "Efficient Storage of Weather Radar Data," *Software-Practice and Experience*, **27**(6), pp. 623–635, June 1997.
- Kruger, A., P.A. Kucera, W.F. Krajewski, and D.A. Short, "TOGA-COARE Shipborne Radar-Rainfall Products," *Water Resources*

Research, **35**(8), pp. 2597-2600, August 1999.

- Mathews, J.G., "Evaluating Data-Compression Algorithms," *Dr. Dobb's Journal*, pp. 50-53, January 1996.
- Moszkowicz, S., G.J. Ciach, and W.F. Krajewski, "Statistical Detection of Anomalous Propagation in Radar Reflectivity Patterns," *Journal of Atmospheric and Oceanic Technology*, **11**(4), pp. 1026-1034, August 1994.
- Rew, R. and G. Davis, "NetCDF: An Interface for Scientific Data Access," *IEEE Computer Graphics and Applications*, pp. 76-82, July 1990.
- Steiner, M., J.A. Smith, S.J. Burges, C.V. Alonso, and R.W. Darden, "Effect of bias adjustment and rain gauge data quality control on radar rainfall estimation," *Water Resources Research*, **35**(8), pp. 2487-2503, August 1999.