

P1.17 RECENT DEVELOPMENTS IN DATA ANALYSIS, QUALITY CONTROL AND DATA BROWSING AT THE NATIONAL WEATHER SERVICE OFFICE OF HYDROLOGY

Jeng-Jong Pan¹, Geoffrey M. Bonnin, Robert L. Mott, and Heather Friedeman²
National Weather Service, Office of Hydrology, Silver Spring, Maryland

1. INTRODUCTION

Data, and the quality of the data, are critical to the hydrologic mission of National Weather Service (NWS). The operational component of the mission is performed at 13 River Forecast Centers (RFC) and approximately 120 Weather Forecast Offices (WFO) at strategic locations across the United States. The NWS Office of Hydrology supports the operational mission by developing, implementing, and maintaining hydrologic models and systems in cooperation with the field offices. The forecast models used by the field offices are developed and calibrated for specific rivers and streams based on historical events. They are conditioned and constrained operationally using current observations and, in the case of operational ensemble forecasts, with historical data as well. Inaccurate, inconsistent, incomplete or insufficient data can cause significant problems in the forecast process and for the forecasters who operate it.

The Office of Hydrology has recently developed several new approaches for data analysis of both operational and historical data, as well as for archiving and retrieving historical data. The data analysis approaches, are targeted at helping users to reduce uncertainties associated with the use of data, and they include; automated double-mass analysis for inconsistency checks, wavelet transform for nonstationary time-series analysis, cluster analysis for the study of regionalization, and robust outlier detection for spatial inconsistency checks. These approaches are being integrated into our operational development and forecast processes.

Our traditional source of historical data has been the major U.S. collection and archiving agencies such as the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center (NCDC), the U.S. Geological Survey (USGS), and the Natural Resources Conservation Service (NRCS). The Office of Hydrology has recently begun capturing and archiving operational data that has been lost in the past, and has also developed tools that provide new ways to view and access historical data and inventories of historical data.

World Wide Web (WWW) forms have been developed as simplified interfaces to traditional data access utilities and a new interactive "browser" has been developed that allows users to query and browse inventories of the historical data as well as view and retrieve the data itself. The ability to query and view the data inventories greatly improves the user's effectiveness in selecting data for specific projects.

2. DATA ANALYSIS

The Office of Hydrology has developed several data analysis tools for various applications. Our goal is to integrate these analysis tools into the operational computing environment in order to achieve efficient data processing in support of in-house research and field operations. Three of the analysis tools are introduced here.

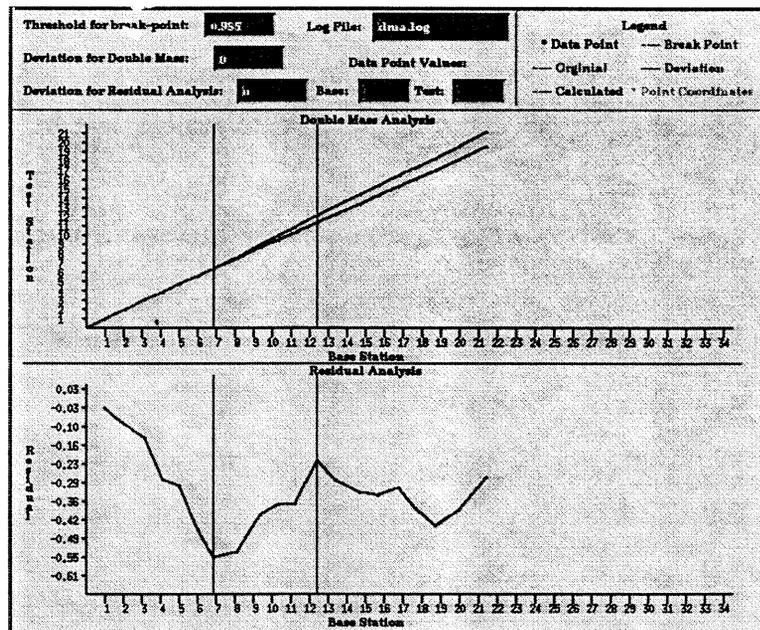


Figure 1. Inconsistency Checks Using DMA Tool

2.1 Automated Double Mass Analysis (DMA) Tool

DMA is a technique commonly employed to detect changes in data-collection procedures or conditions at a given location. The changes may result from changes in instrumentation, changes in observation procedures, or changes in gage location or surrounding conditions. In DMA, a mass curve is a plot of the accumulation of the observed element over time for one location (test station) versus the accumulation over time for a reference location (base station). The mass curve is approximately a straight line if the variations at both test and base stations are quite

¹ Corresponding author address: Jeng-Jong Pan, W/OH1, 1325 East-West Hwy, Silver Spring, MD 20910

² COMSO Inc.

consistent. Any break point in the curve suggests a possible change at the test station in relation to the base station.

The automated DMA tool can detect multiple break points and generate modified data. Figure 1 shows the user interface and the automated initial detection of break points. This GUI-based tool allows users to add or delete break points interactively. Station history information from the data inventory showing when the station has been moved can be integrated into the tool for supporting the user's decisions if a correction is necessary. The automated DMA tool could cut the manual data processing time by 90%. The analysis of variance (Searcy and Hardison, 1960; Chang and Lee, 1974) has been implemented to perform the automated detection. An improvement of break point detection using other mathematical and statistical methods is under study.

2.2 Multiscale Time Series Analysis Tool

To study the multiscale and nonstationary characteristics of a time series, the continuous wavelet transform (CWT) provides the capability to investigate temporal variation with different scales. The CWT is defined as the convolution of a time series $x(t)$ with a wavelet function $\psi(t)$ shifted in time by a translation parameter b and a dilation parameter a (Morlet et al., 1982):

$$S(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt$$

where $*$ is the complex conjugate, and $a (> 0)$ and b are real numbers. The scalogram is defined as $|S(b,a)|^2$. The Morlet wavelet with a constant $c (=5.3)$ used in this tool is adopted here.

$$\psi(t) = e^{ict} e^{-\frac{t^2}{2}}$$

Figure 2 shows a scalogram of a monthly precipitation time series, where wavelet variance (WV) is defined as the integration of the scalogram over time for given scales. Therefore, the wavelet variance is a function of time-scale that represents the marginal density function of energy and shows the relative intensities of a time series at different time-scales. The high magnitude in the scalogram indicates that the monthly precipitation is quasi-stationary at the annual scale (Pan et al., 1997).

2.3 Cluster Analysis Tool

The purpose of cluster analysis of hydrometeorologic stations, also called regionalization, is to decompose a large and climatologically complex area into several smaller climatologically homogeneous regions for quality control and modeling. In the study of regionalization, there are three primary issues: (1) input data, (2) analysis methods, and (3) grouping approaches have critical impacts on the

subregion features. For the data issue, one must consider the data length (i.e., appropriate segmentation of the period of record), data time-scale (i.e., hourly, daily, monthly, etc. values), and data transformation (i.e., conversion to another form, such as square-root, to better identify the variation among individual values). The rotation of principal components (Richman, 1986) has been adopted for cluster analysis. The grouping approach can be either "hard" region (i.e., non-overlapping between regions) or "soft" region (i.e., transition zones could exist between regions).

A special feature of this tool is that users can use the scalogram generated from the multiscale time series analysis tool as input. The regions determined in this way are sensitive to the time-scale (or duration) of the individual data values (Pan et al., 1997). This knowledge allows us to make regionalization choices that are appropriate to the scale of hydrologic applications.

3. QUALITY CONTROL (QC)

The purpose of quality control is to prevent "bad" data from being used in various hydrologic processes (e.g., calibration, modeling, forecasting, etc.). For quality control of hydrometeorologic data one performs a variety of checks, such as range checking, spatial inconsistency checking, temporal inconsistency checking, internal consistency checking, and multi-sensor inconsistency checking (Krajewski, 1986).

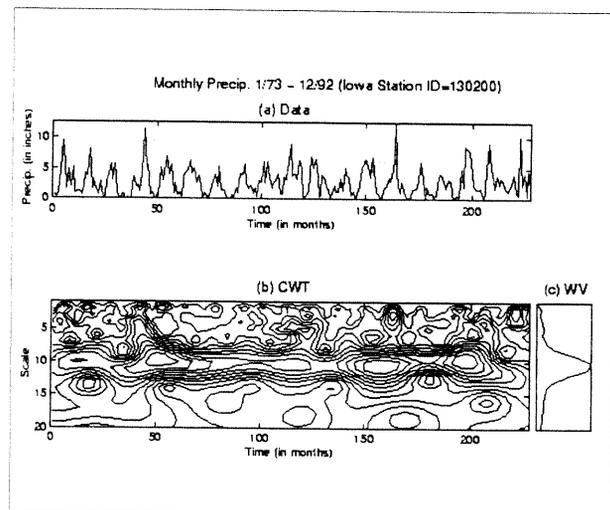


Figure 2. Rainfall Analysis Using CWT

The Office of Hydrology has developed the GUI-based prototype of a real-time quality control system to support field operations. The features of this system include real-time access to operational data as it is captured at the forecast office, efficient and robust outlier detection, reanalysis, integration of a variety of information (e.g., WSR-88D, topo., etc.) for decision making, and multiple temporal scale data handling. Real-time access and processing are important because they support the forecaster's need to

make decisions in short time frames (one minute or less) as the data is arriving. In this paper, we briefly review the strategy and methods used for spatial inconsistency checks.

To satisfy the operational requirements, we have implemented a simple and robust algorithm to perform efficient and effective detection of suspect data, and then we use an estimation method to reanalyze these suspect data. Finally, users can perform multi-sensor data comparison, if possible, to validate these outliers. We can define climate regions for the implementation of spatial inconsistency checks that support efficient real-time quality control. Fovell et. al. (1993) have defined large climate zones of the conterminous United States using monthly precipitation and temperature. The cluster analysis tool mentioned earlier can be applied for regionalization using different temporal scale data as input.

The steps of robust outlier detection are:

Step 1. Determine the median and mean absolute deviation (*MAD*) of *N* stations in each climate region:

$$MAD = \frac{1}{N} \sum_{i=1}^N |x_i - x_{med}|$$

Step 2. Determine *Index1* or *Index2* for each station

i.

if (*MAD* = 0) *Index1*=0

if (*Q*₇₅ ≠ *Q*₂₅) then

$$Index1 = (x_i - Q_{50}) / (Q_{75} - Q_{25})$$

else

$$Index2 = (x_i - Q_{50}) / MAD$$

end if

where *Q_k* is the *k*th percentile (i.e., *Q*₅₀ is the median). Data are temporarily treated as suspect if either *Index1* or *Index2* are greater than predefined thresholds (=2 for both of them). Madsen (1992) has implemented a similar approach for daily precipitation quality control.

In the reanalysis, an estimation method is used to estimate these suspected data. In this prototype the inverse

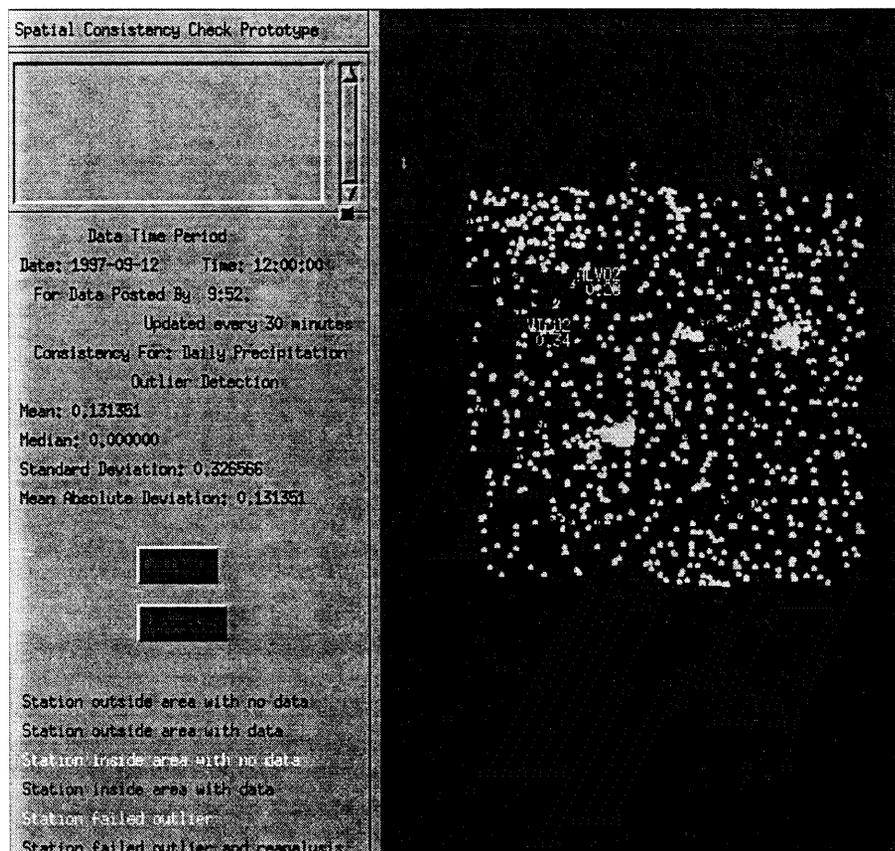


Figure 3. Real Time Quality Control System for Outlier Detection

distance weighting method used in the NWS River Forecast System (NWSRFS) is selected due to its simplicity. Other estimation methods (e.g., Miller et al., 1992) will be integrated into this system to compare their performance. Figure 3 shows the user interface and results of QC. Currently two RFCs are working closely with the Office of Hydrology to test this real-time QC system. The system will also provide several other functions, such as detection of malfunctioning raingages, interactive modification of "bad" data, etc.

4. HISTORICAL ARCHIVES

The Office of Hydrology maintains archives of historical data for use in development and improvement of hydrologic forecasting techniques, and for use in calibrating forecasting techniques to specific basins. The tasks of technique development and calibration are done by both the operational field offices (RFCs) and at the Hydrologic Research Laboratory

4.1 How is the Data Used?

The processes of technique development and model calibration both involve using sets of historical physical elements such as rain, snow, temperature, river stage, river flow, etc. as inputs to systems that allow computed outputs to be compared with observed values. Over the years, the

Office of Hydrology has developed systems used for technique development and model calibration. These systems have well-defined interfaces for data input and as a result, the systems that produce data sets from the historical archives are specifically tailored to produce them in these formats. The systems are therefore designed for efficient use for the particular purposes of the NWS hydrology program. Such specificity brings with it both advantages and disadvantages. The advantages lie in the increased effectiveness of users because of the reduction of effort associated with data handling. The disadvantages lie in the need for development and maintenance of custom software. OH is taking steps to reduce the need for such custom software by storing data in more generic data formats and by exploring ways to reduce the cost of preparing data inventories through cooperation with the data suppliers.

4.2 What Do We Archive

The Office of Hydrology traditionally has used data provided by NCDC (primarily precipitation and other atmospheric elements), USGS (primarily stream flow and stage), and NRCS (primarily snow fall information from their SNOTEL network). These data are archived to fulfill requirements for understanding the nation's natural resources. In some cases the records go back through the nineteenth century. However, there is a significant body of data collected operationally that does not make it into such readily accessible archives. For example, there are many precipitation gages used operationally whose data is not retained as part of any national program. Furthermore, there is a proportion of data collected for understanding the climatic record that is archived but is not gathered sufficiently rapidly to be used operationally. It has been demonstrated (Johnson et. al., 1997; Finnerty et. al., 1997) that the historical and operational data sets are sufficiently different, that calibrations based on historical data produce different results when driven with operational data.

Beginning in 1993, the Office of Hydrology began archiving the data used operationally. Specifically, we began archiving all of the point observations passing through the operational databases of the RFCs as well as retaining their radar precipitation products (WSR-88D Precipitation Processing Subsystem Stage I, II, and III products). Selected subsets of this data are being made more publicly available by the NOAA Joint Office for Science Support using funding from climate research programs such as GCIP (see <http://www.ogp.noaa.gov/gcip/#da>). It is the intent of the Office of Hydrology to logically merge these data with the historical data from the national archiving agencies, thereby improving the information available for calibration and technique development activities.

4.3 Data Inventories

In terms of sheer volume, there is a vast amount of data available in the historical data archives that hydrologists must search through to select data that are appropriate to a particular development activity. In most cases, the data record is not continuous; there are gaps in the record; stations are established, moved, and discontinued; and instruments change as well as do collection times and observers. To improve the effectiveness of the data

The screenshot shows a web form titled "Stations (there are two ways to specify stations)". It contains several sections:

- 1) If you need only a few stations (- 6), enter the state station numbers in the area below (example: 41 7943).** This section has a text input field for "State Station Numbers:".
- 2) Or, if you want many stations (up to 900), make a file that is a list of the numbers.** This section includes instructions on file naming and a text input field for "Enter your file's name that's in 140302D144's pub:".
- Elements:** A list of checkboxes for data elements:
 - PTPX / PRCP - precipitation
 - EPAN / EVAP - pan evaporation
 - SNOW / SNOW - snowfall
 - SNOW / SNWD - point snow depth
 - SNWE / WTEQ - point snow cover/water-equivalent
 - TAMK / TMAX - maximum air temp
 - TAMN / TMIN - minimum air temp
 - TTPMX / MXPN - maximum pan water temp
 - TTPMN / MNPN - minimum pan water temp
 - WUDS / WDMV - point wind travel
- Time boundary choice:** Two input fields for "From month/year" (with "1992" entered) and "through" (with "1994" entered).
- Use:** Radio buttons for "calendar year boundaries (Jan through Dec)" and "water year boundaries (Oct through Sept)".
- Return output as:**
 - A table per station per element per year of your time request
 - A summary table combining time for each station-element request (results are calculated using the bounds of your time request)
 - Generate OH Data cards (suppresses printed table choices above)
 - Pad beginning and ending with missing values if actual values are unavailable
 - Make PDPF Station information headers
 - Make @F @G @H cards for MAT (only works if you selected Temperature elements)
- Units:** Radio buttons for "English units" and "Metric units".

Figure 4. WWW Forms Interface

selection task, the Office of Hydrology has developed a series of inventories for its data sources. The inventories contain such information as data completeness, period of record, and station moves in addition to identity, location, and observed parameters. The maintenance of such compact inventory information allows users to make effective queries without having to search the actual data (in the same manner as one uses a library catalog when searching for information in a library of books). The queries are predictable based on the types of tasks users perform and we have designed and implemented a reduced information layer that satisfies the majority of the data mining queries.

4.4 *Web Forms*

Office of Hydrology historical data formerly resided on batch mainframe computer systems but have been moved to UNIX-based workstations. The original inventory query utilities required punch card input and they produced line-printer output. Users ran a succession of these queries, examining the character-based, tabular results for each run. To provide a transition from the old batch environment to the new interactive environment while new interactive tools were under development, we ported the mainframe utilities to the workstation environment and built WWW forms (Figure 4) as a front end that automatically build the card image input streams. Access to the WWW forms is restricted to specific Internet domains. In other words, we are running an Extranet. We are taking this approach because our mission is focused on specific users and we must be careful with our resources and maintain that focus.

4.5 *Historical Data Browser*

The Historical Data Browser is a new tool that the Office of Hydrology has developed and fielded. It allows users to perform queries against the data inventories, review the results both in geographic and graphical displays, display and plot the actual data, and when satisfied, produce data sets formatted appropriately for subsequent

Figure 5. Historical Data Browser Query Entry

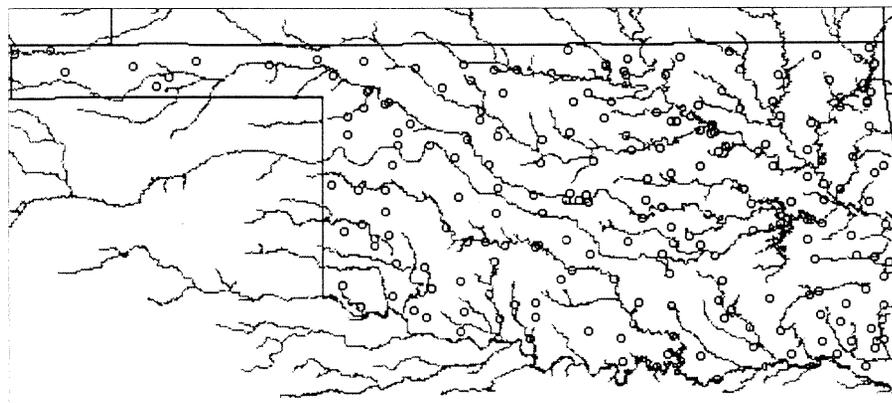


Figure 6. Geographic Display of Stations Satisfying Selection Criteria

use.

In response to a particular query (Figure 5), the Historical Data Browser will display all stations satisfying the specified conditions and display them on a map background as shown in Figure 6.

The user can select stations from the geographic display and show information about the station as well as display the data, if desired, as shown in Figure 7.

The Office of Hydrology maintains the raw historical data centrally, and the Historical Data Browser and the inventory files are distributed to the field offices. This distribution mechanism allows each office to perform inventory queries rapidly at the local site and then retrieve the data from the central repository without the need for replicating the data itself. The Historical Data Browser itself

deals with the physical location of the data and makes its location transparent to the user.

5. CONCLUSION

The NWS Office of Hydrology is introducing a range of new technologies to the problems of operational and historical data analysis, quality control and data management. These new technologies take account of the work flow in the forecast offices and are based in a variety of disciplines from statistics to data management. They show promise in improving the quality and timeliness of hydrologic forecasts as well as making more effective use of the skills of professional employees.

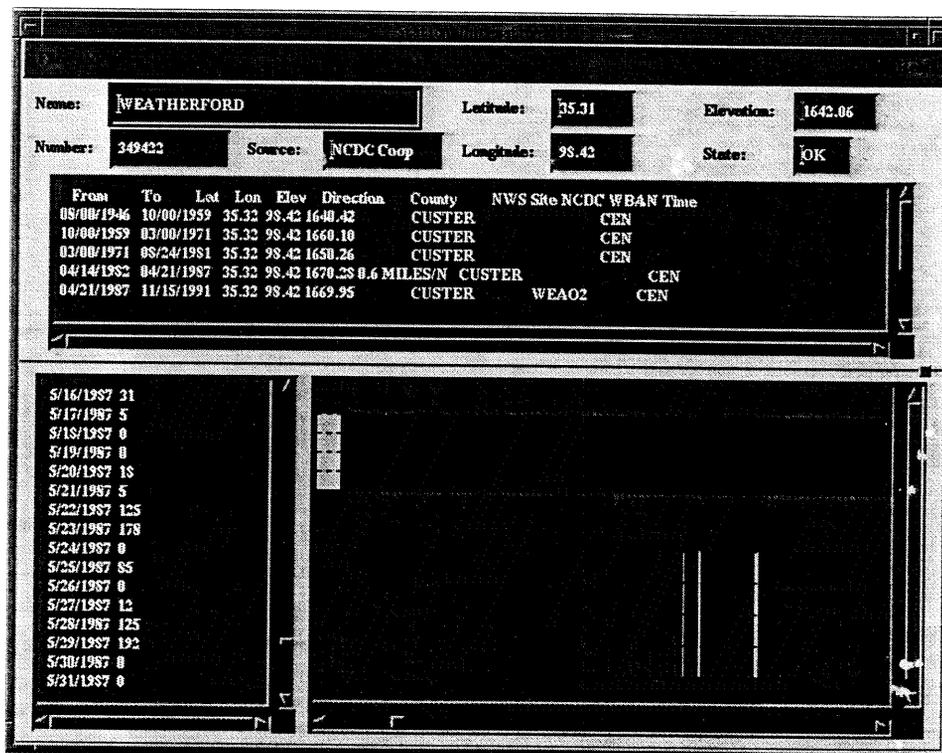


Figure 7. Station Information Display and Data Plot from Historical Data Browser

6. REFERENCES

- Chang, M., and R. Lee, 1974: Objective double-mass analysis, *Water Resources Research*, **10**, 1123-1126.
- Finnerty, B., and D. Johnson, 1997: Comparison of National Weather Service Mean Areal Precipitation Estimates Derived from NEXRAD Radar vs. Raingage Networks, Proceedings of Theme A, *International Association of Hydraulic Research (IAHR) XVII Congress*, San Francisco, Ca. August 10-15, 601-606.
- Fovell, R. G., and M. C. Fovell, 1993: Climate zones of the conterminous United States defined using cluster analysis, *J. Climatol.* **6**, 2103-2135.
- Johnson, D., M. Smith, B. Finnerty, and V. Koren, 1997: Comparing Mean Areal Precipitation Estimates from NEXRAD and Raingage Networks, submitted to the *Journal of Hydrologic Engineering* (in review).
- Krajewski, W. F., 1986: Quality control of hydrometeorological data, *Second Int. Conf. On Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Florida, 112-117.
- Madsen, H., 1992: Semi-automatic quality control daily precipitation measurements, *the 5th Int. Meeting on Stat. Climatology*, Canada, 375-377.
- Miller, P., and S. Benjamin, 1992: A system for the hourly assimilation of surface observations in mountainous and flat terrain, *Monthly Weather Review*, **120**, 2342-2359.
- Morlet, J., G. Arens, I. Fourgeau, and D. Giard, 1982: Wave propagation and sampling theory, *Geophysics*, **47**, 203-236.
- Pan, J. J., S. T. Li, and S. Cong, 1997: Multiscale study of the spatial variability in the cluster analysis of rainfall stations, submitted to the *Int. of J. Climatol.* (in review).
- Richman, M. B., 1986: Rotation of principal components, *J. Climatol.*, **6**, 293-335.
- Searcy, J. K., and C. H. Hardison, 1960: Double-mass curves, *U.S. Geol. Surv. Water Supply Pap.*, **1541-B**, 27-66.