

## QUALITY CONTROL OF HYDROMETEOROLOGICAL DATA

Witold F. Krajewski

Hydrologic Research Laboratory  
National Weather Service, NOAA  
Silver Spring, Maryland

### 1. INTRODUCTION

Quality of hydrometeorological data is a very important factor affecting operations of hydrometeorological services. The real-time forecasting of hydrometeorological processes is especially vulnerable to poor quality data. Dealing with high volumes of data within a limited time framework calls for automated quality control procedures preceding the input of the data to the hydrological and meteorological forecast models. Although there are many aspects of quality control of hydrometeorological data, in this paper we will limit our considerations to those methods and techniques that could be implemented in a fully automated mode in an operational environment.

A uniform approach to quality control of hydrometeorological data is very difficult to develop. One has to face problems that result from sparse networks, high variability in time and/or space of some physical phenomena, wide range of measurement hardware, diverse management of the measurement networks and telecommunication links, and non-standardized structure of the hydrometeorological data bases. Also, factors such as forecast lead time, observation frequency, and the economic implications of wrong forecasts dictate different approaches to data quality control. Yet another consideration is that of computational constraints, important in situations such as flash floods, where prompt and accurate forecasts are of vital interest.

Over the years a number of papers have been published in the literature on the subject of quality control of hydrometeorological data [for example, the works of Reynolds and Campbell (1971), Makhover and Ovsiannikov (1971), Allen (1972), Shearman (1975), and Bryant (1979)]. In this paper we will be interested mainly in those hydrometeorological variables that are useful in real-time hydrologic forecasting. A general conceptual approach to the problem of data quality control in such an environment will be discussed. Of primary interest is precipitation as the main driving force of hydrologic models. Also important are streamflow, air temperature, dewpoint temperature, pressure, wind speed, snow cover, and snow water equivalent. For a more 3

detailed description of the components of common hydrologic models see, for example, Peck (1976).

The occurrence of bad data can be caused by several factors. The ones common to many hydrologic and meteorological variables are: failure, malfunction or improper placement of the sensor or measuring device, outside electromagnetic field interference or malfunction of communication lines, or coding error. There are other causes of erroneous data specific to a particular measurement technology and variable of interest (for example, anomalous propagation in radar measurements of rainfall or navigational error in satellite observations). It should be noted at this point that the objective of quality control is to eliminate gross errors, not the measurement error which is inherent in every technology.

There are at least two possible general approaches to analysis of bad data, a deterministic and a statistical approach. In a deterministic approach, the physical characteristics of the process of interest are used to determine if a data point should be accepted or questioned as bad data. In a statistical approach, the statistical parameters of a variable are used to do the same job. Before we discuss both approaches in more detail, let us distinguish two inherent problems in quality control. The first one is detection of bad data and the second is bad data accommodation or substitution. We will concentrate here on mainly the first problem.

Bad data can be defined as data falling outside some expected range. Thus we will call these data outliers. From that definition, it is clear that an outlier does not necessarily have to be an extreme value, although in many cases it is.

### 2. DETERMINISTIC APPROACH

In a deterministic approach, physical characteristics of a measured process are used in a deterministic way to detect outliers. Usually, that means setting upper and lower limits for the variable of interest. In some cases, those limits are obvious (for example, 0 is a lower limit for precipitation), but in others they are

functions (Hampel, 1974). For any statistic, one can develop an influence function which, roughly speaking, measures the influence of a given data point on that statistic. Ideally, one would like all the points to contribute equally, so if a given point influences the statistic of interest unusually much, it is suspected of being an outlier. For a stationary time series, the primary statistic of interest is the correlation function. The corresponding influence function has the form:

$$I(H, \rho(k), (y_i, y_{i+k})) = y_i y_{i+k} - \frac{1}{2} \rho(k) \cdot (y_i^2 + y_{i+k}^2) \quad (1)$$

where  $y_i, y_{i+k}$  are standardized observations,  $\rho(k)$  is the correlation value of lag  $k$ , and  $H$  is the marginal distribution of  $y_i$ . The distribution of  $I(\cdot)$  has a tractable form and can be used to identify the unusually large values (in absolute terms) at some specified control level. This approach does not require

airplanes. Either case substantially affects the spatial correlation structure of the radar-measured rainfall field. If the radar-rainfall data are given in a digitized form on a rectangular grid (e.g., the HRAP grid, Greene and Hudlow, 1982), then it is very easy to compute the autocorrelation matrix of the data. The next step is to compute the values of the influence function for all the points in the radar field that contributed to the computations of the correlation matrix. The points which belong to many pairs with high influence function values should be questioned. For the details of the computational algorithm, see Krajewski (1985).

The above methods require assumption of stationarity of the time series, an assumption which is often not satisfied in nature. How serious the consequences could be is perhaps best shown by the example given by Subba Rao (1979) in his discussion of a paper by Kleiner et al. (1979). The time series shown in Figure 2 was generated by the process:

$$X_t = 0.4 S_{t-1} + 0.8 X_{t-1} Z_{t-1} + Z_t \quad (2)$$

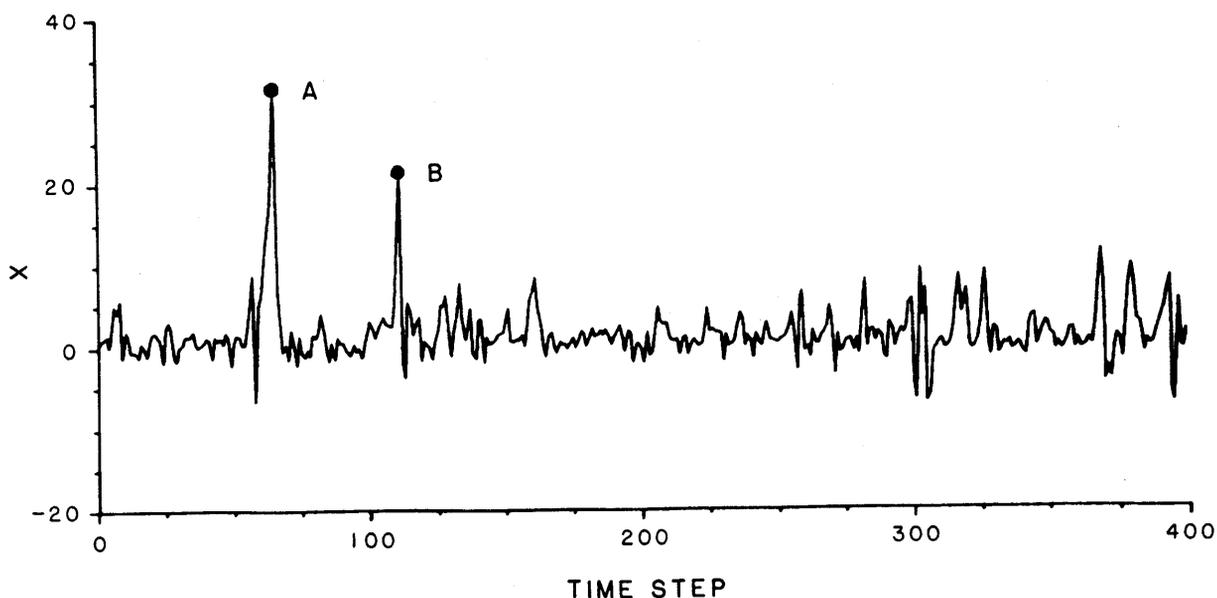


Fig. 2. Nonstationary time series

any model of the time series and could be implemented in real time. For more details, see Chernick et al. (1982).

Similar techniques could be applied to multivariate and univariate samples and also to spatial data. As an example of the latest technique, the concept of the influence function for the correlation function has been applied to quality control of radar-rainfall data. It is a recognized fact that radar-rainfall data, even those obtained from a well calibrated radar, contain a lot of noise and, in some cases, gross errors that result, for example, from anomalous propagation or from reflected echoes of

where the  $Z_t$  are independent standard normal variables. The generated series is nonstationary and does not contain outliers, but if one tried to identify a stationary model for this time series, then the points A and B would be considered as outliers. A significant loss of information would be experienced.

Another group of techniques for outlier detection considers the data as a univariate sample having a certain distribution. The statistical literature is very rich here and many tests and algorithms for parameter and quantile estimation have been devised. An example of such an approach, in the area of hydrometeorological

not easy to find and are determined in a subjective way based on local or global recorded maxima or minima. Another often considered characteristic is rate of change in both space and time. Again, the upper and lower limits are set based on historical data analysis. The third characteristic is the time of no-change. Due to the natural variability and periodicity of most of the hydrometeorological variables, any variable is bound to change after a certain period of time. In general, this period is shorter for values of the variable that are farther from normal. This characteristic is a very useful means of detecting malfunction or failure of the measuring device.

Quality control based on the above characteristics is easy to implement and does not require excessive processing time or computer storage. The danger, however, is the subjectivity in setting the limits. If they are set too high, then the probability of bad data falling into an allowable range is increased. If, on the other hand, the limits are set too low, then the consequences may be even more dramatic, since good but extreme data could be rejected or ignored. Another important point is that exogenous information is not utilized. The decision to accept or reject the data is based solely on analysis of the variable itself. Also, spatial analysis of rate of change is not often performed unless the density of the network is sufficient.

Figure 1 illustrates the deterministic methods of outlier detection. Readers interested in more detailed discussion of implementation aspects should refer to works by Makhover and Ovsyannikov (1971), Van der Schaaf (1984), and O'Brien and Keefer (1985).

### 3. STATISTICAL APPROACH

A discussion of the statistical approach to hydrometeorological data quality control requires additional categorizing of possible situations. From the statistical point of view, hydrometeorological variables can be seen as space and time dependent continuous stochastic processes. These processes are sampled in time and space and collected data constitute discrete spatial time series. Outlier detection in spatial time series is a very little explored, almost nonexistent area, at least as far as formal methods are concerned.

A hydrometeorological variable can be considered simply as a time series, without taking into account its spatial aspects. The few existing formal methods using this approach apply to stationary processes that follow an autoregressive scheme of known order. Certain slowly varying hydrometeorological variables can be modeled that way, especially if the seasonal trends are removed. A maximum likelihood ratio test has been developed by Fox (1972) for such models, and could be performed in real time to check for outliers. It should be pointed out that the model would not be used for forecasting purposes, but only to constitute the framework for outlier detection. The advantage of such an approach is that the limits for maximum and gradient do not need to be specified or known. The disadvantages are that the order and parameters of the autoregressive model need to be estimated and then stored for each station (location).

Another approach to outlier detection in a time series has been suggested by Chernick et al. (1982). It is based on the concept of influence

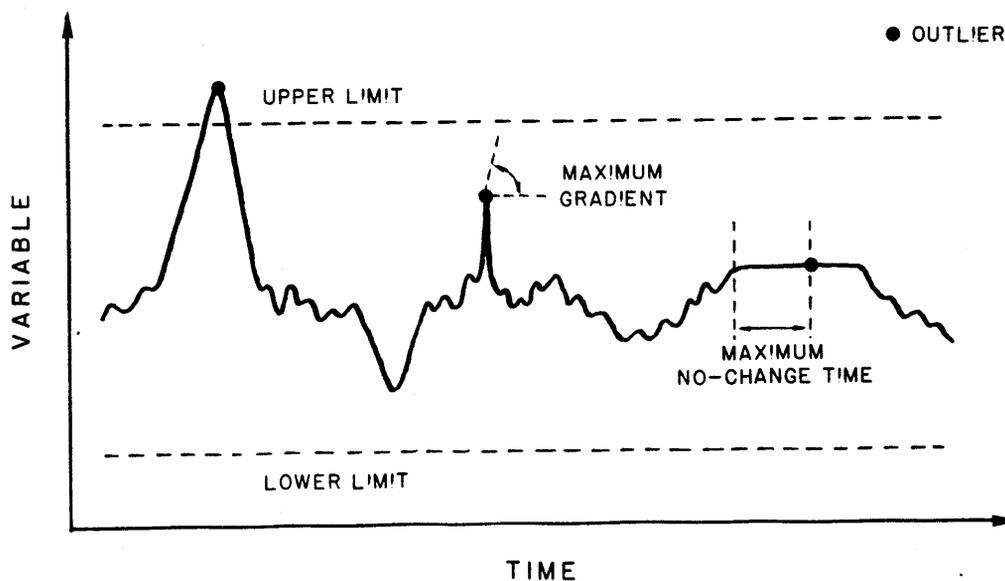


Fig. 1. Illustration of deterministic methods for outlier detection

data quality control, is given in a paper by Bissell (1981).

Bissell applied this approach to screen out bad rain-gage precipitation data. According to his algorithm, the original data values are transformed to values of a standard normal variable through some relationships that account for local, seasonal, and sampling changes. Then, a statistical test is performed to check each data value either against climatological means or forecasted values (if available). This approach requires storage of a set of parameters for each station but is not very demanding computationally.

#### 4. OTHER APPROACHES

In the statistical approach described above, the models were based on data of the variable of interest itself. However, a model can be based on some other variables that are related through some statistical or physical relationship to the variable analyzed. In such cases, a model could be used in a forecasting mode to produce a forecast based on previous data of the variable of interest and current data of related variables. Within this framework it is convenient to quality control the new data through the analysis of residuals of the forecast. A good forecasting model should be characterized by approximately Gaussian, uncorrelated model errors, constituting an especially simple environment for quality control testing.

Also, if a model is given in state-space form, there is feedback, so that new observations can be used to adjust the initial states of the model, allowing production of better quality forecasts, which in turn can be used to quality control new observations. This type of approach could be very useful, especially in streamflow data analysis using stochastic-dynamic routing models (Georgakakos and Bras, 1982).

Yet another approach to quality control of hydrometeorological data, a very natural one, is cross-referencing using data from multiple sensors. This can be best illustrated with precipitation data. Standard raingages, radars, and satellites are used to collect data from which quantitative estimates of precipitation are derived. Although all these sensors utilize different physical principles of measurement and possess different sampling characteristics and error structures, there is a significant amount of independent information in each data set that can be used to detect outliers and other anomalies in other data sets. An example of how to use such information within a quality control framework is given in a paper (Hudlow et al., 1982) on the Precipitation Processing System planned for NEXRAD (Next Generation Weather Radar).

#### 5. CONCLUSIONS

It is clear from the above discussion that the problem of quality control of hydrometeorological data is a complex and difficult one, even if considered separately from technological issues (sensor design, communication links, etc.). It seems that the most proper and at the same time general approach is to build a system that would include all of the above analyzed elements. These elements should be arranged in a hierarchy of decisions (decision tree). In such a decision tree, the simplest methods capable of detecting the most obvious outliers are placed before the more sophisticated methods. At each level, outlier detection is followed by a common element, i.e., outlier accommodation. An illustration of such a quality control tree is given in Figure 3.

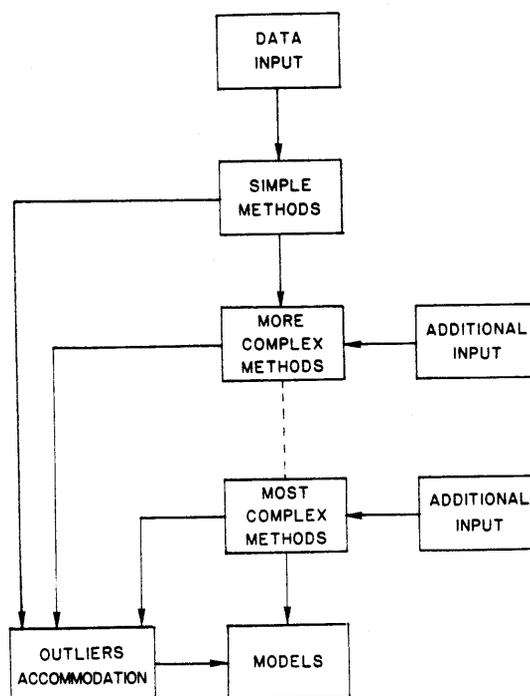


Fig. 3. Quality control decision tree

A slightly different concept could be utilized if a definite decision cannot be made at each level. In that case, it is advisable to assign a probability (or a similar measure of uncertainty) to each decision and proceed to the next (lower) level until the accumulated probability reaches a specified level selected as the rejection threshold. In this approach also, accommodation follows detection.

Finally we present a quality control system for precipitation data, currently being implemented at the Hydrologic Research Laboratory for use with the data available from the NWSRFS (National Weather Service River Forecast System) Version 5 data base (Figure 4). The radar data come from a RADAP II System. The satellite data

are GOES infrared images indicating cloud top temperatures. Inference can be made to determine the regions of precipitation based on the satellite information. This subsequently can help to determine the existence of anomalous propagation in the radar data (Fiore, 1985). Then, the influence function method is used to check for any other outliers. In unclear situations a comparison with the rain gage data will be made after adjustment is made to account for different sampling properties of the two sensors. Rain gage data quality control will be based on historical data distribution parameters and spatial correlation analysis.

#### 6. ACKNOWLEDGEMENTS

The author would like to thank Mrs. Lianne Iseley for assistance with editorial and graphical work, and Drs. Konstantine Georgakakos, Michael Hudlow, Eric Anderson, and Richard Farnsworth for helpful discussions.

#### 7. REFERENCES

Allen, P.G. (1972): The routine processing of current rainfall data by computer. Meteorological Magazine, 101, pp. 340-345.

Bissell, V.C. (1981): Screening techniques for telemetered data in near real-time forecast applications. Proceedings of International Symposium on Real-Time Operation of Hydro-systems, University of Waterloo, Waterloo, Ontario, Canada.

Bryant, G.W. (1979): Archiving and quality control of climatological data. Meteorological Magazine, 108, pp. 309-315.

Chernick, M.R. (1982): The influence function and its application to data validation. American Journal of Mathematical and Management Sciences, Vol. 2, No. 4, 263-288.

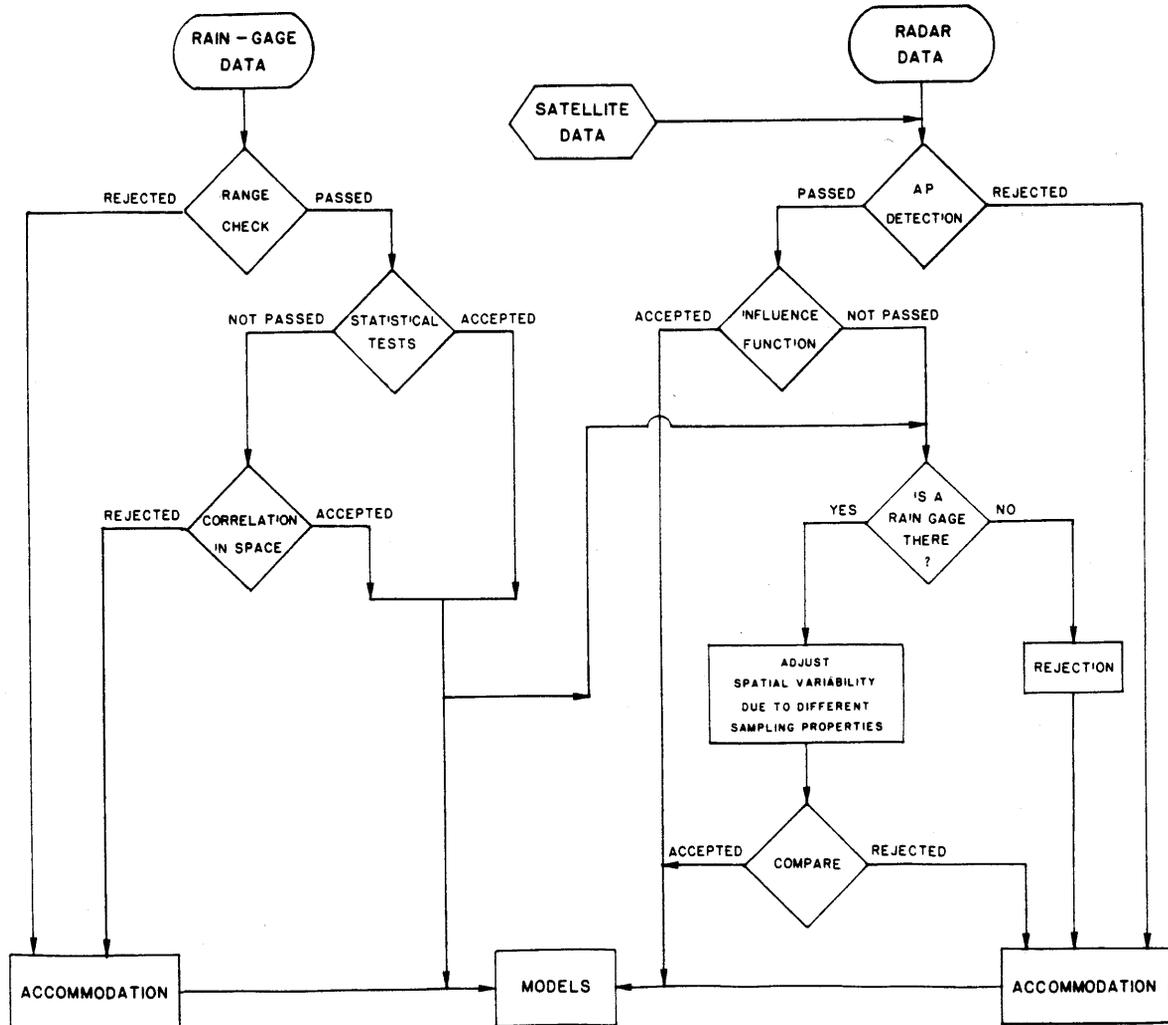


Fig. 4. Precipitation data quality control

- Fiore, J.V. (1985): The use of satellites for rainfall estimation and quality control of radar data. Unpublished report, Department of Meteorology, University of Maryland.
- Fox, A.J. (1972): Outliers in time series. J. Royal Statistical Society, B, 43, pp. 350-363.
- Georgakakos, K.P., and R.L. Bras (1982): Real-time, statistically linearized adaptive flood routing. Water Resources Research, 18(3), pp. 513-524.
- Greene, D.R. and M.D. Hudlow (1982): Hydro-meteorological grid mapping procedures. Unpublished manuscript, Hydrologic Research Laboratory, NWS, NOAA, Silver Spring, Maryland.
- Hampel, F.R. (1974): The influence curve and its role in robust estimation. Journal of American Statistical Association, 69, pp. 383-393.
- Hudlow, M.D., D.R. Greene, P.R. Ahnert, W.F. Krajewski, T.R. Sivaramakrishnan, E.R. Johnson, and M.R. Dias (1982): Proposed off-site precipitation processing system for NEXRAD. Preprints of 21st Conference on Radar Meteorology, Edmonton, Alberta, Canada, pp. 394-403.
- Kleiner, B., R.D. Martin, and D.J. Thompson (1979): Robust estimation of power spectra. J. Royal Statistical Society, B, 41, pp. 313-351.
- Krajewski, W.F. (1985): Quality control of radar-rainfall data using influence functions. To be submitted to the Journal of Atmospheric and Oceanic Technology.
- Makhover, Z.M. and V.V. Ovsyannikov (1971): Automatic monitoring and processing of hourly hydrometeorological information. Proceedings of the International Symposium of Hydrometeorologists from the Socialist Countries, translated from Russian by the Israel Program for Scientific Translations.
- O'Brien, K.J. and T.N. Keefer (1985): Real-time data verification. Preprints of ASCE Conference on Computer Applications in Water Resources, Buffalo, New York.
- Peck, E.L. (1976): Catchment modeling and initial parameter estimation for the National Weather Service River Forecast System. NOAA Technical Memorandum NWS HYDRO-31, U.S. Dept. of Commerce, Silver Spring, Maryland.
- Reynolds, G.W. and R.H. Campbell (1971): A computerized method of telemetered precipitation data quality control. Journal of Weather Modification, 3(1), pp. 235-243.
- Shearman, R.J. (1975): Computer quality control of daily and monthly rainfall data. Meteorological Magazine, 104, pp. 102-108.
- Subba Rao, T. (1979). Discussion of paper by B. Kleiner, R.D. Martin, and D.J. Thompson (1979). J. Royal Statistical Soc., B, 41, pp. 346-347.
- Van der Schaaf, S. (1984): Errors in level recorder data--prevention and detection. Journal of Hydrology, 73, pp. 373-382.