

# Radar Rainfall Data Quality Control by the Influence Function Method

WITOLD F. KRAJEWSKI

*Hydrologic Research Laboratory, National Weather Service, National Oceanic and Atmospheric Administration, Silver Spring, Maryland*

The statistical concept of the influence function is applied to detection of outliers in radar rainfall fields. The method can identify observations which are inconsistent with the spatial correlation in the field. A Monte Carlo experiment has been performed to test the method for daily and hourly radar rainfall data and to compare it to other simple methods such as range and gradient checks. The results of that study indicate the usefulness of the method in the detection of outliers in real time. They also show the potential of the method to deal with outliers resulting from certain types of anomalous propagation.

## 1. INTRODUCTION

The ability of a weather radar to monitor rainfall continuously in time and space is very attractive from an operational point of view. As a result, there are numerous examples of the application of radar for hydrologic purposes [e.g., *Kessler and Wilk*, 1968; *Anderl et al.*, 1976; *Collier et al.*, 1983; M. D. Hudlow, unpublished manuscript, 1973]. Systems have been developed that are capable of producing digitized estimates of rainfall with high temporal and spatial resolution (e.g., the Radar Data Processor II system; D. Greene et al., unpublished manuscript, 1983). However, there are some problems with a direct use of radar rainfall estimates as an input to hydrologic or other types of operational models. Although radar provides a very good areal description of rainfall coverage, the magnitude of rainfall estimates is contaminated by errors of a multiple nature. A good discussion of various sources of error in radar rainfall estimates is given in the work by *Harrold et al.* [1973] and *Wilson and Brandes* [1979]. The errors can be reduced if radar rainfall estimates are combined with other rainfall data such as rain gage or possibly satellite data. Papers by *Eddy* [1979], *Crawford*, [1979], and *Krajewski and Hudlow* [1983] address the problem of optimal merging of radar and rain gage rainfall data. Such procedures are, however, very sensitive to the quality of radar data and the density and configuration of rain gage networks. Thus in order to provide high-quality radar rainfall data, additional processing must precede the merging procedures (M. D. Hudlow et al., unpublished manuscript, 1983; P. R. Ahnert et al., unpublished manuscript, 1983). This paper describes a preprocessing procedure applicable to this quality control problem.

High-quality rainfall data are understood to be data which do not contain unrealistic values hereinafter called "outliers." Outliers in radar rainfall data can result from anomalous propagation (AP), interference of signal from other than rain targets such as airplanes or towers, communication line problems, etc. Detection of outliers in radar rainfall data seems, on the surface, to be a simpler problem than it really is. The crux of the problem is to develop a usable definition of "unrealistic." For instance, let's consider high values. If we set an arbitrary

upper limit on "realistic" values, we risk that if the limit is set too low, we will reject correct values and underestimate the rainfall. Or, if the limit is too high, some incorrect high values or outliers can slip through the system and cause overestimation of rainfall. In either case, underestimation or overestimation of rainfall can significantly affect the subsequent analysis. Trying to set the limit based on historical data analysis is difficult, since the result would depend upon the geographic location, the availability of the historic data, and the approach taken. In the case of radar rainfall data, the historic data base is very limited and because of different sampling and error characteristics cannot be substituted or even supplemented by a rain gage data base. Similar problems exist for low values. Although the physical lower limit on radar rainfall values is simply zero, it is possible that unusually low data values appear in a region of intense precipitation. In this case, it is even more difficult to establish a simple threshold. Similar considerations apply to analysis of gradients instead of magnitudes in the radar rainfall field. It is clear from the above discussion that simple threshold tests are inadequate and that other approaches are necessary.

In this paper, application of the influence function method to the problem of outlier detection in radar rainfall fields is described. The method is attractive since it can be fully automated and used in an on-line mode. Also, it does not require any external information or calibration. The method, as applied to radar rainfall fields, is based on analysis of spatial correlation in the field. Outliers are defined as data being inconsistent with the correlation function. Such definition allows for detection of the high-valued outliers as well as the low-valued points. The assumptions required are that the rainfall field is statistically homogenous (up to second-order moments) and can be transformed into a Gaussian field. It seems that both assumptions are satisfied in most nonorographic type situations. A more detailed description of the influence function concept is given next.

## 2. MATHEMATICAL BACKGROUND

### 2.1. Influence Function Method

Rainfall is considered as a stochastic process; this paper is focused on statistical analysis of samples from this stochastic process. In any analysis of statistical data, one faces a problem of robust estimation, i.e., constructing such estimates that would be resistant to differences in realizations (or samples) of the underlying process. A related problem is that of estimation

TABLE 1. Critical Probability Values

a	$I_{cr}$					
	1.00	2.00	3.00	4.00	5.00	6.00
1.00	0.2090	0.0618	0.0196	0.0065	0.0022	0.0007
0.90	0.1814	0.0477	0.0135	0.0040	0.0012	0.0004
0.80	0.1525	0.0346	0.0085	0.0022	0.0006	0.0002
0.70	0.1224	0.0231	0.0047	0.0010	0.0002	0.0000
0.60	0.0918	0.0136	0.0022	0.0004	0.0001	0.0000
0.50	0.0618	0.0065	0.0007	0.0001	0.0000	0.0000
0.40	0.0346	0.0022	0.0002	0.0000	0.0000	0.0000
0.30	0.0135	0.0004	0.0000	0.0000	0.0000	0.0000
0.20	0.0022	0.0000	0.0000	0.0000	0.0000	0.0000
0.10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Constant  $a = 1 - \rho^2(k, l)$ .

in the presence of outliers. *Hampel* [1974] introduced a concept of influence curve or influence function to facilitate analysis of contaminant effects on estimators. In that concept, all data obey a basic model (or distribution)  $F$ . The data contain a contaminant which comes from a different model  $G$ . Both models constitute a contamination model  $C$ :

$$C = (1 - \lambda)F + \lambda G \tag{1}$$

where  $\lambda \in \langle 0, 1 \rangle$ . Now, let's denote our estimator of interest by  $T$  (for instance,  $T$  can be the mean, the variance, the correlation coefficient, etc.) and our data sample by  $Z = \langle Z_1, Z_2, \dots, Z_n \rangle$ . Of course,  $T = T(Z_1, Z_2, \dots, Z_n)$ ; i.e.,  $T$  is a function of  $Z$ .

In order to evaluate the influence of any data point  $\xi$  of the sample  $Z$  ( $\xi = Z_i, i = 1, 2, \dots, n$ ) on the estimator  $T$  we define the influence function

$$f_{T,F}(\xi) = \lim_{\lambda \rightarrow 0} \{ [T((1 - \lambda)F + \lambda G) - T(F)] / \lambda \} \tag{2}$$

We note that

$$f_{T,F}(\xi) = \frac{\partial}{\partial \lambda} \{ T[(1 - \lambda)F + \lambda G] \}_{\lambda=0} \tag{3}$$

$\rho(-2,2)$		.....		$\rho(2,2)$
⋮		$\rho(0,0)$		⋮
$\rho(-2,-2)$		.....		$\rho(2,-2)$

Fig. 1. Spatial correlation matrix.

2	2	2	2	2
2	2	2	2	2
2	2	48	2	2
2	2	2	2	2
2	2	2	2	2

Fig. 2. Critical level exceedance count pattern for an outlier.

*Devlin et al.* [1975] give expressions for the influence function for various estimators, including correlation coefficient in a bivariate sample.

Ideally, each data point should influence the estimator  $T$  to approximately the same degree. If the point  $\xi$  influences  $T$  much more than other points, it can be suspected of being an outlier. The final determination of whether it is an outlier or not is based on the test of influence function value at some specified critical level.

2.2. Influence Function of Spatial Correlation

In this paper we examine the applicability of the influence function concept to the detection of outliers in the rainfall radar field when the estimator of interest is the spatial correlation. Using the result employed by *Devlin et al.* [1975], the influence function for correlation in space is

$$I[H, \rho(k, l), (y_{ij}, y_{i+k,j+l})] = y_{ij}y_{i+k,j+l} - \frac{1}{2} \rho(k, l)(y_{ij}^2 + y_{i+k,j+l}^2) \tag{4}$$

2	2			
2	2	2		
2	2	28		
2	2	2		
2	2	2	2	

Fig. 3a. Pattern along the radar umbrella bound.

2	2	2	2	2
4	2	8	2	2
6	6	48	2	2
4	4	2	2	2
4	4	4	2	2

Fig. 3b. Pattern in a high-gradient region.

where  $I(\cdot)$  is the influence function;  $H$  is the marginal distribution of standard normal observation  $y_{ij}$ ; and  $\rho(k, l)$  is the correlation at lags  $k$  and  $l$  in the  $i$  and  $j$  directions, respectively. Since  $I(\cdot)$  is a function of sample values, it is a statistic itself and can be described by its distribution. It is difficult to derive the distribution of  $I(\cdot)$  directly from (4); therefore a transformation is convenient [see Chernick et al., 1982]:

$$U_{ij}^{kl} = \frac{1}{2}[(y_{ij} + y_{i+k,j+l})(1 + \rho(k, l))^{-1/2} + (y_{ij} - y_{i+k,j+l})(1 - \rho(k, l))^{-1/2}] \quad (5)$$

$$V_{ij}^{kl} = \frac{1}{2}[(y_{ij} + y_{i+k,j+l})(1 + \rho(k, l))^{-1/2} - (y_{ij} - y_{i+k,j+l})(1 - \rho(k, l))^{-1/2}] \quad (6)$$

Then

$$I_{ij}^{kl} = I[H, \rho(k, l), (y_{ij}, y_{i+k,j+l})] = [1 - \rho^2(k, l)]U_{ij}^{kl}V_{ij}^{kl} \quad (7)$$

It can be easily shown that for a stationary Gaussian field,  $U_{ij}^{kl}$  and  $V_{ij}^{kl}$  are independent and normal with a zero mean and variance of one,  $N(0, 1)$ . Thus  $I_{ij}^{kl}$  has distribution of a constant times a product of the standard Gaussian independent variables. It can be shown that  $I_{ij}^{kl}$  has the following probability density function:

$$P(I_{ij}^{kl}) = [1 - \rho^2(k, l)]^{-1} \frac{1}{\pi} K_0(z) \quad (8)$$

where  $K_0$  is the Bessel function of the first kind of order zero, and  $z$  is the product of  $U_{ij}^{kl}$  and  $V_{ij}^{kl}$ .

On the basis of this distribution, a critical value  $I_{cr}$  of  $I_{ij}^{kl}$  can be selected. Table 1 shows the critical probabilities as a function of  $I_{cr}$  and the constant  $[1 - \rho^2(k, l)]$ . It is clear from the table that if one selects  $I_{cr}$  to be 4.0, then the probability associated with it is less than 0.01 no matter what value the correlation takes. The sensitivity of the method to the choice of the parameter  $I_{cr}$  is investigated in the following chapters.

As can be seen from (4), in the case of correlation influence function we are actually examining a pair of observations  $y_{ij}$

and  $y_{i+k,j+l}$ . In general, it is difficult to distinguish which point from that pair should be questioned. However, this could be determined if both points are given another test, this time entered coupled with different points. Repeated high value of influence function for a given point usually means the point is an outlier.

### 3. DETECTION OF OUTLIERS IN RADAR RAINFALL DATA

In actual applications of the method, we do not know the true statistics, and therefore estimates have to be used. These will be denoted hereinafter by a circumflex. Let us assume that radar rainfall data are given on a rectangular grid and denote  $R(i, j)$ ,  $i = 1, \dots, NX$  and  $j = 1, \dots, NY$ . A correlation matrix of radar data for nonzero areas within a radar umbrella can be computed from the following formula:

$$\hat{\rho}(k, l) = \frac{1}{N\hat{\sigma}^2} \sum_{m_i}^{M_i} \sum_{m_j}^{M_j} [R(i, j) - \hat{\mu}][R(i + k, j + l) - \hat{\mu}] \quad (9)$$

where  $N$  is the number of pairs of observations whose coordinates differ by the vector  $(k, l)$  and  $R(i, j) > 0$  and  $R[i + k, j + l] > 0$ ,  $\hat{\mu}$  is the estimate of the mean of the nonzero part of radar rainfall field, and  $\hat{\sigma}^2$  is the estimate of the corresponding variance.

Integration limits  $m_i, m_j$ , and  $M_i, M_j$  for a rectangular field can be computed as

$$m_i = \max(1, 1 - k) \quad m_j = \max(1, 1 - l) \\ M_i = \min(Nx, Nx - k) \quad M_j = \min(NY, NY - l)$$

For example, if  $k \in \langle -2, 2 \rangle$  and  $l \in \langle -2, 2 \rangle$ , then the correlation matrix is a  $5 \times 5$  matrix of the form presented in Figure 1.

Suppose now that at the location  $(i, j)$  in our radar rainfall field there is an outlier. If we compute the influence function for all the pairs in the field that are separated by no more than two lags in any direction, and we count, for each pair, the number of times the influence function exceeds the critical level, then the display of these counts for the vicinity of location  $(i, j)$  would result in the pattern shown in Figure 2. Thus the problem of outlier detection in radar rainfall fields can be simplified to the problem of pattern recognition. Of course, a single pattern such as the one presented in Figure 2 appears only if the outlier is very distinct. A more difficult to recognize pattern can result along the boundaries of the search region or in the presence of high local gradients in the radar rainfall field. Figure 3 shows examples of such situations.

Also, it is obvious that the choice of the critical level of influence function affects the pattern as well; this is demonstrated in Figure 4. One can see that the pattern obtained for the critical level equal to 3.0 is easier to recognize than the patterns for  $I_{cr} = 1.0$  and  $I_{cr} = 2.0$ .

The following rules were employed to recognize an outlier pattern.

1. All the locations within a given number of lags from the point under examination must indicate critical level exceedance. The number of lags should correspond to the dimensions of the correlation matrix to be computed.

TABLE 2a. GATE Daily Data Characteristics

GATE Day No.	Field Characteristics				
	$\hat{\mu}$ , mm/h	$\hat{\sigma}$ , mm/h	$R_{max}$ , mm/h	$VR_m$ , mm h <sup>-1</sup> lag <sup>-1</sup>	$\hat{\rho}(1)$
242	0.59	0.57	3.34	2.35	0.95
243	0.08	0.17	2.66	1.96	0.81
244	0.03	0.08	1.41	1.39	0.64
245	1.00	0.85	5.62	3.40	0.90
246	0.11	0.21	3.98	3.14	0.84
247	0.79	1.18	7.07	4.01	0.97
248	1.04	0.95	8.91	8.32	0.91
249	0.28	0.45	4.46	4.20	0.90
250	0.18	0.22	2.81	2.04	0.81
251	0.30	0.55	5.30	3.31	0.92
252	0.55	0.59	5.30	3.28	0.87
253	0.11	0.32	5.62	4.04	0.82
254	0.18	0.40	5.01	3.27	0.84
255	0.61	0.80	5.95	3.34	0.94
256	0.71	0.94	14.12	9.39	0.86
257	1.00	1.02	7.49	5.12	0.92
258	0.27	0.51	3.98	2.95	0.92
259	1.02	0.80	9.99	8.50	0.80
260	0.55	0.65	6.30	2.96	0.90
261	0.11	0.18	3.16	2.53	0.70

$I_{CR} = 3.0$

2	2	2	2	2
2	2	2	2	2
2	2	48	2	2
2	2	2	2	2
2	2	2	2	2

$I_{CR} = 2.0$

4	4	2	2	2
4	4	2	2	2
6	2	48	2	2
6	4	4	2	2
4	4	4	2	4

$I_{CR} = 1.0$

10	4	2	2	2
10	8	4	2	2
10	8	6	48	2
10	8	4	8	8
10	8	8	4	10

Fig. 4. Critical level effect on outlier pattern.

2. The count for a point under examination must be at least four times higher than the average count in the surrounding area (5 × 5 in our examples).

These rules are purely empirical. The second rule could be replaced by a more theoretically sound approach of influence function for the local average of counts, but it was found to be unnecessary in our applications. Further work with real-time radar rainfall data is required for a more thorough examination of these rules. It may be possible to relate the rules to the

TABLE 2b. GATE Day 245 Hourly Data Characteristics

GATE Day 245, hour	Field Characteristics				
	$\hat{\mu}$ , mm/h	$\hat{\sigma}$ , mm/h	$R_{max}$ , mm/h	$VR_m$ , mm h <sup>-1</sup> lag <sup>-1</sup>	$\hat{\rho}(1)$
0100	0.21	0.94	22.38	19.40	0.58
0200	0.34	1.51	28.18	19.77	0.73
0300	0.48	2.44	58.08	35.30	0.78
0400	0.38	1.59	35.48	20.91	0.77
0500	0.31	1.41	42.16	40.67	0.64
0600	0.30	1.15	28.18	22.56	0.65
0700	0.34	1.29	23.71	21.20	0.63
0800	0.58	2.14	56.23	39.45	0.74
0900	0.60	2.09	50.11	34.27	0.74
1000	0.46	1.56	31.62	25.67	0.68
1100	0.61	1.88	29.85	25.87	0.69
1200	0.86	2.28	33.49	28.18	0.71
1300	1.19	3.05	47.31	36.58	0.74
1400	1.62	3.37	42.16	28.04	0.77
1500	1.78	3.61	37.58	26.61	0.60
1600	1.70	3.37	47.31	33.51	0.80
1700	2.10	3.92	44.66	40.20	0.81
1800	2.15	3.67	39.81	38.69	0.87
1900	2.15	3.80	39.81	38.14	0.89
2000	1.85	3.55	28.18	18.04	0.92
2100	1.66	3.40	35.48	19.21	0.93
2200	1.19	2.52	28.18	22.13	0.88
2300	0.83	2.20	56.23	32.52	0.78
2400	0.48	1.31	22.38	12.94	0.80

TABLE 3a. Outliers Generated for Daily Data

GATE Day No.	Generated Outliers, mm/h					
	1	2	3	4	5	6
242	0.76	6.66	2.52	3.63	7.91	18.26
243	1.07	0.82	0.92	1.04	0.83	1.85
244	1.07	0.97	1.26	0.83	0.93	1.11
245	15.48	1.97	1.54	14.68	0.50	4.77
246	2.17	1.34	1.58	1.74	2.13	1.36
247	0.61	11.27	9.42	9.20	0.02	2.51
248	0.45	22.75	0.97	0.26	4.66	40.88
249	0.46	1.29	7.27	0.99	2.42	0.26
250	1.07	0.81	0.97	1.37	1.38	1.45
251	3.93	1.66	0.27	0.14	2.88	4.32
252	1.21	5.03	4.17	0.17	2.47	5.21
253	0.93	3.58	1.81	4.91	0.91	1.13
254	0.90	2.32	0.61	1.12	1.25	1.88
255	0.91	0.53	12.92	11.67	9.61	0.49
256	0.45	0.34	0.44	0.55	6.30	5.55
257	1.04	0.82	2.18	17.57	0.71	8.23
258	0.36	2.22	3.90	0.41	0.90	1.19
259	4.72	4.95	11.64	2.57	7.99	8.04
260	1.17	5.62	4.56	0.13	2.56	5.84
261	1.01	2.12	1.45	2.51	1.00	1.12

TABLE 4a. Number of Outliers Detected by Simple Methods, Daily Data

Gate Day No.	Maximum Range Method	Maximum Gradient Method
242	1	3
243	0	0
244	0	0
245	2	2
246	0	0
247	1	3
248	1	3
249	0	1
250	0	0
251	0	0
252	0	2
253	0	0
254	0	0
255	1	3
256	0	1
257	1	2
258	0	0
259	1	3
260	0	2
261	0	0

physical size of rain cells. In order to evaluate the performance of the above described method, a Monte Carlo experiment was performed and is described next.

4. MONTE CARLO EXPERIMENT

Evaluation of the performance of the influence function method and a sensitivity analysis of the critical level choice

was carried out through a simulation-type experiment. Daily and hourly radar rainfall data from the international GARP Atlantic Tropical Experiment (GATE) conducted in 1974 were used. The choice of GATE data was motivated by the high quality of the data which was assured by a very thorough post experiment data processing and analysis. It is highly unlikely that any outliers exist in that data set. This is an important point, since our performance test was based on analysis of

TABLE 3b. Outliers Generated for Hourly Data

GATE Day 245, hour	Generated Outliers, mm/h					
	1	2	3	4	5	6
0100	0.42	5.02	5.29	0.36	3.90	0.16
0200	8.94	0.84	0.38	0.49	65.11	2.78
0300	0.00	0.00	0.03	0.02	25.18	0.03
0400	0.10	0.05	0.87	4.71	0.02	0.83
0500	1064.31	0.86	2.62	0.58	2.01	0.98
0600	0.42	0.21	2.03	3.54	35.87	4.98
0700	2.11	1.83	1.50	7.00	0.03	0.01
0800	0.11	1.31	0.04	0.00	0.21	31.40
0900	0.75	0.00	0.00	0.98	3.28	751.02
1000	265.88	6.65	0.51	0.20	5.36	2.36
1100	0.35	0.09	183.03	143.98	91.24	0.08
1200	0.20	0.51	0.01	0.23	4.39	0.01
1300	0.44	0.01	0.00	0.00	1.39	0.00
1400	0.01	0.20	0.42	1.64	0.01	0.56
1500	0.00	0.89	59.69	3.25	2382.37	14.00
1600	0.00	8.06	6.43	0.00	0.00	0.00
1700	0.05	6.86	0.54	0.00	1.23	21.16
1800	0.00	70.68	0.00	0.14	0.02	177.44
1900	0.11	0.00	35.77	227.00	47.25	4.02
2000	0.00	2.60	7971.37	0.00	0.00	2325.89
2100	1.04	5704.68	0.01	0.00	0.00	0.08
2200	1.72	0.00	0.00	129.11	11.67	0.02
2300	184.04	2.15	4.25	18.81	7.91	686.77
2400	2.26	0.57	1.91	30.57	46.62	0.00

TABLE 4b. Number of Outliers Detected by Simple Methods, Hourly Data

Gate Day No.	Maximum Range Method	Maximum Gradient Method
0100	0	0
0200	1	1
0300	0	0
0400	1	0
0500	0	1
0600	0	1
0700	0	0
0800	0	0
0900	1	1
1000	1	1
1100	3	3
1200	0	0
1300	0	0
1400	0	0
1500	2	2
1600	0	0
1700	0	0
1800	2	2
1900	1	3
2000	2	2
2100	1	1
2200	1	1
2300	2	2
2400	0	1

TABLE 5a. Number of Outliers Detected, Daily Data

GATE Day No.	$I_{cr}$											
	1.0		2.0		3.0		4.0		5.0		6.0	
	A	B	A	B	A	B	A	B	A	B	A	B
242	4	4	4	4	4	4	4	4	4	4	4	4
243	6	6	6	6	2	2	1	1	1	1	1	1
244	4	4	7	5	0	0	0	0	0	0	0	0
245	2	2	3	3	3	3	3	3	3	3	3	3
246	4	4	5	5	5	5	5	5	5	5	5	5
247	3	3	4	4	4	4	4	3	3	3	3	3
248	9	4	5	3	4	2	3	2	3	2	3	2
249	7	4	4	4	3	3	2	2	2	2	1	1
250	4	3	2	2	2	2	2	2	1	1	1	1
251	5	4	4	4	3	3	3	3	3	3	3	3
252	6	4	5	4	4	3	4	3	4	3	2	2
253	9	5	10	5	10	6	6	3	2	2	2	2
254	4	1	1	1	1	1	1	1	1	1	1	1
255	6	3	5	3	3	3	3	3	3	3	3	3
256	1	1	1	1	1	1	1	0	0	0	0	0
257	1	1	2	2	2	2	2	2	2	2	2	2
258	4	2	4	1	1	0	1	0	1	0	1	0
259	3	2	2	2	3	3	3	3	1	1	1	1
260	3	3	3	3	3	3	3	3	3	3	3	3
261	5	3	4	3	2	2	1	1	1	1	1	1

Column A contains the number of detected outliers, and column B contains the number of actual outliers.

TABLE 5b. Number of Outliers Detected, Hourly Data

GATE Day No.	$I_{cr}$											
	1.0		2.0		3.0		4.0		5.0		6.0	
	A	B	A	B	A	B	A	B	A	B	A	B
0100	0	0	0	0	0	0	0	0	0	0	0	0
0200	2	2	1	1	1	1	1	1	1	1	1	1
0300	2	1	2	1	1	1	1	1	1	0	0	0
0400	0	0	0	0	0	0	0	0	0	0	0	0
0500	1	1	1	1	1	1	1	1	1	1	1	1
0600	3	1	1	1	1	1	1	1	1	1	1	1
0700	0	0	0	0	0	0	0	0	0	0	0	0
0800	1	1	1	1	1	1	1	0	0	0	0	0
0900	2	1	1	1	1	1	1	1	1	1	1	1
1000	5	2	1	1	1	1	1	1	1	1	1	1
1100	4	2	2	2	3	3	3	3	3	3	2	2
1200	0	0	0	0	0	0	0	0	0	0	0	0
1300	1	0	0	0	0	0	0	0	0	0	0	0
1400	1	0	0	0	0	0	0	0	0	0	0	0
1500	1	1	1	1	1	1	1	2	2	2	2	2
1600	1	1	0	0	0	0	0	0	0	0	0	0
1700	0	0	0	0	0	0	0	0	0	0	0	0
1800	2	2	2	2	2	2	2	2	2	2	2	2
1900	2	2	3	2	2	2	2	2	2	2	2	2
2000	5	2	3	2	2	2	2	2	2	2	2	2
2100	1	1	3	2	2	1	2	1	2	1	1	1
2200	4	2	3	2	2	2	2	2	1	1	1	1
2300	3	3	3	3	3	3	3	3	3	3	3	3
2400	3	2	2	2	2	2	2	2	2	2	2	2

Column A contains the number of detected outliers, and column B contains the number of actual outliers.

generated outliers. Six outliers were generated for each field from a lognormal distribution with mean equal to  $\hat{\mu}$  and high variance equal to 100. That way, some of the outliers were clearly outside of any reasonable physical range, some were high by these standards but realizable, and some were as low as most of the values in the field and thus almost impossible to detect.

Some of the characteristics of 20 selected GATE days, as well as those from each of the 24 hours of GATE day 245 are summarized in Tables 2a and 2b, respectively. The selection of the 20 days was done arbitrarily. In Table 2,  $\hat{\mu}$  denotes the mean value of nonzero region in the field,  $\hat{\sigma}$  is the corresponding standard deviation,  $R_{max}$  is the maximum value in the field,  $\nabla R_m$  is the maximum gradient in the field, and  $\hat{\rho}(1)$  is lag-one correlation coefficient averaged over N-S and E-W directions. Table 3 includes the generated values of outliers, six for each field. Table 4 presents the results of application of two simple methods of outlier detection. One method is called maximum range check and the second is maximum gradient check. For the purpose of this study, the maximum range value was selected as 10.00 mm/h for daily data (Table 4a) and 50 mm/h for hourly data (Table 4b). The maximum gradient was set as 5 mm h<sup>-1</sup> lag<sup>-1</sup> for daily and 30 mm h<sup>-1</sup> lag<sup>-1</sup> for hourly data. These decisions are arbitrary and were made without a long preceding study, but here they serve an illustrative purpose only. Table 5 contains the results of the influence function method application for different critical values. One can see that when the critical value equals 1.0, the method is over-sensitive and detects relatively high values in the fields as the outliers. Raising the control level to 2.0 helps to avoid this problem, and raising it to 3.0 virtually eliminates it. Some-

times, for a low critical level, even high-valued outliers are undetected because the critical level exceedance pattern becomes too messy and does not conform to the detection rules. Raising the initial level clears the picture and the outliers can be detected.

For a more convenient evaluation of performance of the influence function method, three additional performance statistics were computed (R. J. Donaldson et al., unpublished manuscript, 1975) as follows: (1) probability of detection, defined as the ratio of correctly detected outliers to the total number of outliers; (2) false alarm ratio, defined as the ratio of incorrectly detected outliers to the total number of detections; and (3) critical success index, defined as the ratio of correctly detected outliers to the sum of the total number of outliers and incorrectly detected outliers. The results averaged over all days and all hours are given in Tables 6 and 7, respectively.

Comparison of the results in Tables 4 and 5 clearly favors the influence function method (for example, compare day 246 daily rainfall values), which is geared to individual fields and defines outliers through correlation, a statistic very important in any analysis of radar rainfall fields. The results in Tables 6 and 7 should be used only in a relative sense. They do not indicate the absolute performance measures of the methods. As we pointed out earlier, the generation scheme used in this study affects heavily the "configuration" (or distribution) of outliers and biases the results. This can be easily seen in hourly data where only a small part of all generated outliers deserves such a name; however, in the computation of the summary statistics the total number was used.

TABLE 6. Summary Statistics for Daily Data

	Critical Level $I_{cr}$						Maximum Range	Maximum Gradient
	1.0	2.0	3.0	4.0	5.0	6.0		
Probability of detection	0.525	0.542	0.433	0.383	0.333	0.317	0.067	0.208
False alarm	0.234	0.140	0.108	0.104	0.079	0.067	0.001	0.011
Critical success index	0.429	0.478	0.406	0.365	0.325	0.311	0.067	0.202

In operational application of the method, it is suggested that a prespecified value for the critical level be used instead of expressing it in terms of critical probability, since this would require computation of an appropriate quantile for each single value in the field and thus prolong the processing. Based on the results of this study, it is recommended that  $I_{cr} = 3.0$  be used.

The existence of outliers can significantly affect the estimates of the correlation matrix. Therefore the influence function method should be applied in two or more passes. Often, in the first pass only the most apparent outliers are detected, since their existence has a masking effect on other outliers. The detected outliers should be accommodated, i.e., substituted by some other values we think are the best estimates of clearly wrong values. The concept of the influence function method allows for a simple method for outlier accommodation. They should be replaced with values which would not affect the estimates of the statistic of interest (the correlation in this case) computed without them. In our study, we did not apply this approach in a strict sense. We substituted for the outliers the local averages; a procedure which has little influence on the correlation in the field. The averages were computed from the eight surrounding values.

## 5. CONCLUSIONS AND FINAL REMARKS

The influence function method and its application for the quality control of radar rainfall data has been presented. The results of a Monte Carlo experiment indicate the usefulness of the method. The main assumptions of the method, as applied to the radar rainfall field, are that the fields are Gaussian and second-order stationary. Both assumptions, although in gener-

al not met by rainfall fields, are approximately met if a normalizing transformation is applied and rainfall accumulated data are considered. In case of strong orographic effects, the methods can be used to analyze the residuals from the mean (if such can be identified and estimated). In principle, the method can also be used for nonhomogenous fields. It would require, however, a derivation of influence function for generalized covariance (see, for example, *Bras and Rodrigues-Iturbe* [1985]). Applicability of the method to analyze radar data from localized thunderstorms seems to be even more limited. Some practical applications are needed to fully explore this problem.

The power of the influence function method used to quality control radar rainfall data lies in the fact that it does not require knowledge of the distribution of the data nor of any other characteristics that cannot be derived from the available data. It does not require the setting of arbitrary limits on the magnitude or gradient in the rainfall field and it accounts for local anomalies. This means that the outliers need to be abnormal in a local sense only in order to be detected. Also, the method is capable of detecting a "negative" outlier, i.e., a value that is much smaller than its neighbors (but not necessarily negative). Other advantages of the method are its simplicity to implement, a calibration-free operation with a sensitivity tuning capability, and statistical soundness.

The method can also handle some types of AP existence. If AP appears in a rainy region, it is often manifested as high-gradient, high-magnitude data. It would obviously affect the correlation structure of the radar rainfall field. Clear air AP usually cannot be detected by the influence function method but then other methods such as comparison with satellite data (J. V. Fiore et al., unpublished manuscript, 1986) or hardware-type methods can be used [*Aoyagi*, 1983].

TABLE 7. Summary Statistics for Hourly Data

	Critical Level $I_{cr}$						Maximum Range	Maximum Gradient
	1.0	2.0	3.0	4.0	5.0	6.0		
Probability of detection	0.186	0.174	0.174	0.174	0.160	0.153	0.125	0.152
False alarm	0.258	0.076	0.021	0.021	0.021	0.000	0.118	0.288
Critical success index	0.167	0.168	0.172	0.172	0.159	0.152	0.122	0.142

*Acknowledgments.* The author gratefully acknowledges the comments and suggestions of E. R. Johnson of the Hydrologic Research Laboratory of the National Weather Service and K. P. Georgakakos of the University of Iowa. The editorial work of R. Ripkin is also appreciated.

## REFERENCES

- Anderl, B., W. Attmannspacher, and G. A. Schultz, Accuracy of reservoir inflow forecasts based on radar rainfall measurements, *Water Resour. Res.*, 12(2), 217-223, 1976.
- Aoyagi, J., A study on the MTI weather radar system for rejecting ground clutter, *Pap. Meteorol. Geophys.*, 33(4), 187-243, 1983.
- Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, Mass., 1985.
- Chernick, M. R., D. T. Downing, and D. H. Pike, Detecting outliers in time series data, *J. Am. Stat. Assoc.*, 77, 743-747, 1982.
- Collier, C. G., P. R. Larke, and B. R. May, A weather radar correction procedure for real-time estimation of surface rainfall, *Q. J. R. Meteorol. Soc.*, 109, 589-608, 1983.
- Crawford, K. C., Considerations for the design of a hydrologic data network using multivariate sensors, *Water Resour. Res.*, 15(6), 1752-1762, 1979.
- Devlin, S. T., R. Gnanadesikan, and T. R. Kettenring, Robust estimation and outlier detection with correlation coefficients, *Biometrika*, 62, 531-545, 1975.
- Eddy, A., Objective analysis of convective scale rainfall using gages and radar, *J. Hydrol.*, 44, 125-134, 1979.
- Hampel, R. F., The influence curve and its role in robust estimation, *J. Am. Stat. Assoc.*, 67, 383-393, 1974.
- Harrold, T. W., E. J. English, and C. A. Nicholas, The Dee weather radar project: The measurements of area precipitation using radar, *Weather*, 28, 332-338, 1973.
- Hudlow, M. D., and V. L. Patterson, GATE Radar Rainfall Atlas, NOAA special report, 155 pp., U.S. Dep. of Commerce, Washington, D. C.
- Kessler, E., and K. E. Wilk, Radar measurement of precipitation for hydrological purposes, *WMO/IHD Rep. 5*, World Meteorol. Org. and Int. Hydrol. Decade, Geneva, 1968.
- Krajewski, W. F., and M. D. Hudlow, Evaluation and application of a real-time method to estimate mean areal precipitation from gage and radar data, paper presented at the WMO Conference on Mitigation of Natural Hazards, World Meteorol. Org., Sacramento, Calif.
- Wilson, T. W., and E. A. Brandes, Radar measurement of rainfall—A summary, *Bull. Am. Meteorol. Soc.*, 60(9), 1048-1058, 1979.

W. F. Krajewski, Hydrologic Research Laboratory, National Weather Service, National Oceanic and Atmospheric Administration, Silver Spring, MD 20910.

(Received April 28, 1986;  
revised January 5, 1986;  
accepted January 6, 1986.)