

# Multivariate Short-Term Rainfall Prediction

EDWARD R. JOHNSON

*National Weather Service, Hydrologic Research Laboratory, Silver Spring, Maryland 20910*

RAFAEL L. BRAS

*Department of Civil Engineering, Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139*

Accurate hydrologic forecasting for basins with a short response time depends on the ability to predict quantitative rainfall rates. This study develops a stochastic model for short-term (of the order of 1 hour or less) rainfall prediction. The model simultaneously predicts rainfall rates at multiple locations and for multiple values of prediction lead. All model parameters are estimated solely from telemetered rain gage data for the event being predicted. The model includes velocity and direction of storm movement as explicit parameters. The storm arrival time at each predicted point is likewise an explicit parameter, which is estimated for each location. The mean rainfall rate is not modeled as being either homogeneous spatially or stationary (constant with time). Likewise, the variance of rainfall is nonhomogeneous and nonstationary.

## INTRODUCTION

In increasing numbers, urban centers are implementing sophisticated real-time control systems for their combined and storm sewers [Labadie *et al.*, 1975]. The ability of these control systems to respond quickly to the characteristics of a precipitation event increases the value of a short-term rainfall prediction. In this context, Grigg *et al.* [1974] have stated, 'Real-time control requires that estimates of interior and overall storm parameters be made. These estimates may be updated as the storm progresses in time as actual data becomes available and their uncertainty will thus decrease. However, the uncertainty will never reach zero until the particular event is over, thus the best level of performance obtained is directly related to storm prediction capability.'

Ideally, the control system should take advantage of the temporal and spatial variability of rainfall/runoff to allocate its hydraulic resources (for example, storage, flow capacity, and treatment capacity) in an optimal fashion. For example, it should be possible to identify the optimal time to use limited storage capacity to avoid overflows or local flooding. It should be possible to trade upstream and downstream resources, for example, to use upstream storage to decrease the load on an overloaded downstream line. This kind of control behavior is dependent on the ability to anticipate future flows with spatial and temporal detail.

In this paper a model for quantitative short-term rainfall prediction that could be used in this and other applications is described. Predictions are made at multiple points to provide some degree of spatial detail and at multiple values of time lead, to provide a degree of temporal detail. The model is nonstationary in structure. The need for a nonstationary rainfall-forecasting model has been previously recognized by Jamieson and Wilkinson [1972].

Besides nonstationarity a successful rainfall prediction scheme requires (1) assumptions about the structure of rainfall, (2) a mathematical model consistent with the assumed structure, and (3) a procedure to estimate the parameters of the mathematical model. The task is to develop a model which has a rich enough structure to reproduce the important

elements of rainfall but is simple enough to allow parameter estimation.

## CHOICE OF MODEL

Throughout this paper, rainfall is modeled as a nonstationary process. In particular, the mean and variance of rainfall are not temporally or spatially constant.

The obvious disadvantage of a nonstationary model is the need to describe and estimate the parameters of the nonstationarity. Postponing for the moment the parameter estimation problem, a nonstationary multivariate model of rainfall is developed. Let

$$i(t) = m(t) + r(t) \quad (1)$$

where

- $i(t)$  vector of  $N$  rainfall rates at time step  $t$  at locations  $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ ;
- $m(t)$  vector of mean values at time step  $t$ ;
- $r(t)$  vector of residuals at time step  $t$ ;
- $N$  number of rain gages in the prediction scheme.

It is notationally convenient to define a diagonal standard deviation matrix  $\Sigma(t)$  with the form

$$\Sigma(t) = \begin{bmatrix} \sigma(x_1, y_1, t) & 0 & 0 & 0 \\ 0 & \sigma(x_2, y_2, t) & 0 & 0 \\ 0 & \vdots & \vdots & \vdots \\ 0 & 0 \dots & \sigma(x_N, y_N, t) & \end{bmatrix}$$

Using this device

$$r(t) = \Sigma(t)\epsilon(t) \quad (2)$$

where  $\epsilon(t)$  is a zero mean, unit variance, random vector process.

The heart of any rainfall prediction scheme is the assumed form of the dynamic behavior of rainfall—how it evolves in time. From (1) the evolution of rainfall in time has two components, changes in the mean value and changes in the residual.

The mean value vector  $\mathbf{m}(t)$  is assumed to change with time in a deterministic, but unknown, fashion. The method proposed to estimate the time-varying mean vector for present and future times will be discussed later. Assume, for the moment, that both the mean vector  $\mathbf{m}(t)$  and the standard deviation matrix  $\Sigma(t)$  are known for all time steps.

At issue in this section is the dynamics of the residual term  $\mathbf{r}(t)$ . The residual will be assumed to evolve in time according to a nonstationary Markov model of the form

$$\mathbf{r}(t + \tau) = A(t, \tau) \mathbf{r}(t) + B(t, \tau) \mathbf{w}(t, \tau) \quad (3)$$

where

- $A(t, \tau)$   $N \times N$  state transition matrix at time step  $t$  for a transition  $\tau$  steps into the future;
- $\mathbf{w}(t, \tau)$   $N \times 1$  vector of disturbances with zero mean value;
- $B(t, \tau)$   $N \times N$  matrix giving the effect of the noise terms at time step  $t$  on the residuals at time step  $t + \tau$ .

Equation (3) is somewhat unusual for including a dependence of  $A$ ,  $B$ , and  $\mathbf{w}$  on the lead value  $\tau$ . This is necessary because of the intention to produce rainfall predictions at multiple values of lead.

At each time step, measured values of rainfall rate are available at a number of points where rain gages are located. The  $N$  points at which predictions are to be made form a subset of the entire rain gage network. For these prediction points a measurement equation is written

$$\mathbf{z}(t) = \mathbf{q}(t) - \mathbf{m}(t) = \mathbf{r}(t) + \mathbf{v}(t) \quad (4)$$

where

- $\mathbf{q}(t)$   $N \times 1$  vector of observed rainfall;
- $\mathbf{m}(t)$   $N \times 1$  vector of mean values;
- $\mathbf{r}(t)$   $N \times 1$  vector of true values of residual;
- $\mathbf{v}(t)$   $N \times 1$  vector of measurement errors;
- $\mathbf{z}(t)$   $N \times 1$  vector of measured residuals;
- $N$  number of points predicted.

It is assumed that the state noise is uncorrelated in time and is uncorrelated with the measurement noise and the measurement noise is uncorrelated in time.

Equations (3) and (4) form the classic framework for the discrete Kalman filter. The derivation is available in many standard texts on estimation [e.g., *Sage and Melsa*, 1971; *Gelb*, 1974] and will not be repeated here. The only complication in the present case is the consideration of multiple-lead predictions. The filter equations are written for a one-step lead as

$$\hat{\mathbf{r}}(t + 1|t) = A(t, 1) \hat{\mathbf{r}}(t|t) \quad (5)$$

$$P(t + 1|t) = A(t, 1)P(t|t)A^T(t, 1) + B(t, 1)E[\mathbf{w}(t, 1)\mathbf{w}^T(t, 1)]$$

$$\cdot B^T(t, 1) = A(t, 1)P(t|t)A^T(t, 1) + Q(t, 1) \quad (6)$$

$$K(t) = P(t|t - 1) \{P(t|t - 1) + E[\mathbf{v}(t)\mathbf{v}^T(t)]\}^{-1} \quad (7)$$

$$\hat{\mathbf{r}}(t|t) = \hat{\mathbf{r}}(t|t - 1) + K(t) \{\mathbf{z}(t) - \hat{\mathbf{r}}(t|t - 1)\} \quad (8)$$

$$P(t|t) = \{I - K(t)\} P(t|t - 1) \quad (9)$$

where  $\hat{\mathbf{r}}(t_2|t_1)$  denotes the linear minimum variance estimate of the true residual vector  $\mathbf{r}(t_2)$ , based on all information available up to time step  $t_1$ .

And

$$P(t_2|t_1) = E\{(\hat{\mathbf{r}}(t_2|t_1) - \mathbf{r}(t_2))(\hat{\mathbf{r}}(t_2|t_1) - \mathbf{r}(t_2))^T\} \quad (10)$$

In words,  $P(t_2|t_1)$  is the error covariance matrix for the estimate of  $\mathbf{r}(t_2)$  made at time step  $t_1$ .  $K(t)$  is called the Kalman gain matrix. To simplify notation, the  $Q(t, \tau)$  matrix has been defined as

$$Q(t, \tau) = B(t, \tau)E[\mathbf{w}(t, \tau)\mathbf{w}^T(t, \tau)]B^T(t, \tau) \quad (11)$$

It is the  $Q(t, \tau)$  matrix and not  $B(t, \tau)$  alone which is important to the prediction.

In addition, it is necessary to define starting conditions  $P(0|0)$  and  $\mathbf{r}(0|0)$ . This is reasonably straightforward. The initial state estimate is taken as the observation before the beginning of rainfall. Therefore the initial estimate of the residual is  $\hat{\mathbf{r}}(0|0) = \mathbf{0}$ , leading to an initial state error equal to the measurement error, i.e.,  $P(0|0) = E[\mathbf{v}(0)\mathbf{v}^T(0)]$ .

Equations (5)–(9) provide the minimum variance estimate of the current residual (equation (8)), the error covariance matrix of the estimate of the current residual (equation (9)), the minimum variance linear prediction of the residual one time step from now (equation (5)), and the error covariance matrix of the one-step prediction (equation (6)). The equations operate recursively; that is, they process only one time step of measurements at a time.

Predictions at the other leads require two new equations to be introduced

$$\hat{\mathbf{r}}(t + \tau|t) = A(t, \tau) \hat{\mathbf{r}}(t|t) \quad (12)$$

$$P(t + \tau|t) = A(t, \tau)P(t|t)A^T(t, \tau) + Q(t, \tau) \quad (13)$$

The rainfall prediction at any future time  $t + \tau$  is given by

$$\hat{\mathbf{i}}(t + \tau|t) = \mathbf{m}(t + \tau) + \hat{\mathbf{r}}(t + \tau|t) \quad (14)$$

The error covariance matrix of this prediction is given by  $P(t + \tau|t)$  (assuming that  $\mathbf{m}(t + \tau)$  is known with certainty).

The quantity  $\{\mathbf{z}(t) - \hat{\mathbf{r}}(t|t - 1)\}$  in (8) is called the innovation vector—it is the difference between the measured residual and its predicted value.

Expanding the innovation expression,

$$\{\mathbf{z}(t) - \hat{\mathbf{r}}(t|t - 1)\} = \{\mathbf{q}(t) - \mathbf{m}(t) - \hat{\mathbf{r}}(t|t - 1)\} \quad (15)$$

The above depends on the time-varying mean vector  $\mathbf{m}(t)$ . As will be seen later,  $\mathbf{m}(t)$  is estimated exogenously to the filtering algorithm. The innovation (equation (15)) is then approximated by

$$\mathbf{q}(t) - \hat{\mathbf{m}}(t|t) - \hat{\mathbf{r}}(t|t - 1)$$

Operationally, no error will be associated with the mean estimate  $\hat{\mathbf{m}}(t|t)$ . As will be seen, dimensionality, uncertain dynamics, and variable states prevent inclusion of the unknown mean among the state variables.

#### PARAMETER REQUIREMENTS

In order to implement the prediction scheme it is necessary to estimate two matrices describing the dynamics of the rainfall residuals at each time step and prediction lead,  $A(t, \tau)$  and  $Q(t, \tau)$ . First, recall the definition of  $Q(t, \tau)$  (equation (11)) and also the state dynamic equation (equation (3)).

The assumption that  $E[\mathbf{w}(t, \tau)] = \mathbf{0}$  has been stated previously. It is necessary to add here that  $E[\mathbf{w}(t, \tau) \mathbf{r}^T(t)] = \mathbf{0}$ .

By postmultiplying both sides of (3) by  $\mathbf{r}^T(t)$  and taking expected values of the resulting equation it is easy to show that  $A(t, \tau)$  is given by

$$A(t, \tau) = \Sigma(t + \tau)E[\mathbf{e}(t + \tau)\mathbf{e}^T(t)](E[\mathbf{e}(t)\mathbf{e}^T(t)])^{-1} \Sigma(t)^{-1} \quad (16)$$

Introducing the notation

$$D(t_1, t_2) = E[\epsilon(t_1)\epsilon^T(t_2)] \quad (17)$$

for the covariance of the normalized residuals allows (16) to be written compactly as

$$A(t, \tau) = \Sigma(t + \tau)D(t + \tau, t)D(t, t)^{-1}\Sigma(t)^{-1} \quad (18)$$

Equation (18) is the framework for estimation of  $A(t, \tau)$ , provided  $\Sigma(t)^{-1}$  and  $D(t, t)^{-1}$  exist. The procedure used to estimate  $D(t, t)$  guarantees the existence of the inverse, as will be briefly seen later. The  $\Sigma(t)$  is a diagonal matrix with diagonal terms equal to the standard deviation of the elements of the residual vector  $r(t)$ . The difficulty is that some element of  $r(t)$  might have zero variance. In this case, the corresponding column of  $A(t, \tau)$  will be taken to be a unit vector, i.e., all zeroes except a value of 1 on the diagonal of  $A(t, \tau)$ . For example, when  $\Sigma(t)$  is a zero matrix,  $A(t, \tau)$  is an identity matrix.

When can rainfall have a variance of zero? One reasonable answer is that rainfall has a mean value of zero and a variance of zero when it is surely not raining. The following assumptions are made for the sake of a reasonable view of rainfall: (1) When the mean value of rainfall is zero, the variance is also zero and (2) when the variance of rainfall is zero, the mean value is also zero. Using these two assumptions, whenever an element of  $r(t)$  has zero variance, that element is identically zero. Notice that the above argument applies to the model of rainfall 'reality.' It is not in contradiction to the initial estimate assumptions given in the previous section.

Postmultiplying (3) by itself transposed, taking expectations, and solving for  $Q(t, \tau)$  yield

$$Q(t, \tau) = \Sigma(t + \tau)D(t + \tau, t + \tau)\Sigma^T(t + \tau) - A(t, \tau)\Sigma(t)D(t, t)\Sigma^T(t)A^T(t, \tau) \quad (19)$$

Substituting (18) for  $A(t, \tau)$  and dropping the transpose from symmetric matrices give

$$Q(t, \tau) = \Sigma(t + \tau) \{ D(t + \tau, t + \tau) - D(t + \tau, t)D(t, t)^{-1}D^T(t + \tau, t) \} \Sigma(t + \tau) \quad (20)$$

Equation (20) provides the framework for estimation of  $Q(t, \tau)$ .

In order to implement the rainfall prediction method described above it is necessary to know the following: (1) measured rainfall at each prediction point at the current time step ( $q(t)$ ), (2) the mean value of rainfall at each prediction point for the current time step ( $m(t)$ ) and for all predicted time steps ( $m(t + \tau)$ ), (3) The standard deviation of rainfall at each prediction point for the current time step ( $\sigma(t)$ ) and for all predicted time steps ( $\sigma(t + \tau)$ ), (4) the measurement error covariance matrix  $E[v(t)v^T(t)]$ , and (5) the covariance matrix of the normalized residuals at each prediction point for the current time step with itself ( $D(t, t)$ ), for each predicted time step with itself ( $D(t + \tau, t + \tau)$ ), and for each predicted time step with the current time step ( $D(t + \tau, t)$ ).

Estimation and prediction of the mean and variance will be discussed in some detail later in the paper. Obtaining the normalized residuals covariance matrix at various lags is a major and important step. The covariance is a function of storm velocity. Johnson and Bras [1978b] discuss the details of all the estimation problems and compare the performance of several methodologies. This paper will briefly present the selected procedure.

### Mean and Variance Estimation

Besides a physical model of rainfall there are two basic statistical approaches to estimating the nonstationary mean and variance required by the model. A multirealization approach assumes that the current rainfall event is a member of a class of rainfall events all with the same mean and variance function. Taking a multirealization approach requires a scheme for classifying events and identifying the event class of each particular event a priori for prediction purposes. Any parameter of the mean and variance function of each class of events would also have to be estimated a priori for each event.

A single-realization approach assumes that each event is unique. As a result no scheme is needed to classify events. This is an advantage for two reasons: there is no need to process historical data and there is no prediction error introduced by misclassification.

The technique used to estimate the nonstationary mean and variance of rainfall is a single-realization approach called the storm counter method. It depends on one fundamental assumption: in relation to the time that a storm arrives at a location the time history of the mean and variance is identical for all locations. It is not assumed that the storm will arrive at all locations, for example, convective cells might completely miss some points.

Data from rain gages are recorded in discrete time steps. The storm counter records the number of time steps since the start of rainfall at each rain gage. The process of assigning storm counters is illustrated in Figure 1. The clock time is 3:00 P.M. Data are collected every 12 min. At 3:00 P.M. it has rained at three rain gages (numbers 4, 6, and 9). Rainfall began at 9:36 A.M. at gage 6, so gage 6 has a storm counter of 27. The storm counter for gage 4 is 15; for gage 9, it is 21.

The mean and variance can be estimated for each storm counter using the common statistical expression for the sample mean and variance:

$$\hat{m}(s) = \frac{1}{n(s)} \sum_{\substack{i=1 \\ t(i,s) \neq 0}}^M q_i(t(i,s)) \quad (21)$$

where

- $\hat{m}(s)$  estimate of mean rainfall rate for storm counter  $s$ ;
- $n(s)$  number of gages with storm counter at least as high as  $s$ , i.e., sample size for storm counter  $s$ ;
- $t(i, s)$  indexing function giving time step for storm counter  $s$  at gage  $i$  (If storm counter  $s$  at gage  $i$  is not available,  $t(i, s) = 0$ );
- $q_i(t)$  measured rainfall rate at gage  $i$  at time step  $t$ ;
- $M$  total number of gages in rain gage network.

$$\hat{\sigma}^2(s) = \frac{1}{n(s) - 1} \sum_{\substack{i=1 \\ t(i,s) \neq 0}}^M [q_i(t(i,s)) - \hat{m}(s)]^2 \quad (22)$$

where  $\hat{\sigma}(s)$  is the estimated variance of rainfall rate for storm counter  $s$ .

It is acknowledged that the variance estimated by (22) is biased, since  $q(t)$  includes an uncorrelated sampling error with given variance. In relative terms this bias is minor and is certainly within the uncertainty of the estimation of the mean. Compensating the result for the known measurement error variance was not considered a worthwhile exercise.

The sample size  $n(s)$  may vary considerably from one storm counter to another. In the case shown in Figure 1 the sample size is 3 for each storm counter from 1 through 15, 2 for storm

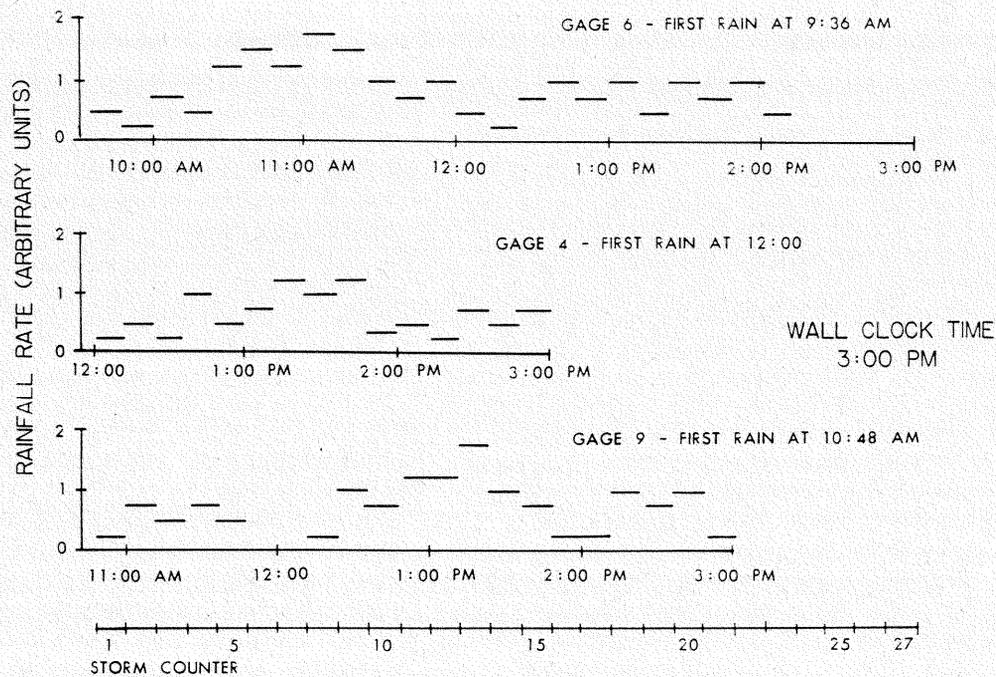


Fig. 1. Illustration of storm counter method.

counters 16 through 21, 1 for storm counters 22 through 27, and no sample is available at all for storm counters 28 or higher.

The storm counter method is easy to explain and to implement. For the prediction problem it is hoped that it will begin to rain first at gages outside the prediction area so that a sample will be available to predict the mean at the prediction points. For the case shown in Figure 1 a sample size of 2 is available to estimate and predict the mean and variance of rainfall at gage 4 for the next 6 time steps (time steps 28–33, storm counters 16–21).

Certain effects are ignored by the storm counter method. For example, the storm counter method does not consider orographic effects or storm aging (i.e., variations of the storm in real time). The basic idea in using a nonstationary mean is that the mean function will capture some of the structure of the event. Ignoring some structural details places a heavier burden on the stochastic components of the prediction scheme. By ignoring orographic effects or storm aging the storm counter method produces higher estimates of the nonstationary variance than a model which accurately includes these effects. However, an inaccurate orographic model might have higher variance than the simple storm counter method, likewise for an inaccurate model of any type of storm structure. Simultaneous occurrence of more than one storm over the area cannot be satisfactorily handled by the methodology.

The storm counter method will produce a mean function that captures more of the structure of some types of events (for example, a frontal storm) than of others (for example, stationary cells). Likewise, the prediction scheme as a whole will perform better for some types of events than for others. The important point is that the storm counter method does not require a particular type of event or impose a particular spatial structure.

Estimation of the unknown nonstationary mean function within the Kalman filter algorithm is not possible. State augmentation which includes the mean (together with rainfall

and residual) will lead to an unobservable system. It is impossible to distinguish between mean and residual when the available observation is their sum [Gelb, 1974, pp. 69–70]. State augmentation would also lead to essentially a 3-fold increase in dimension of the state vector. When forecasts are needed at several stations, the numerical problem would be tremendous. In particular, notice that the storm counter method utilizes all stations in computing the nonstationary mean (and variance). The state-space formulation only includes the gages to be forecasted. The authors are not familiar with any way to formulate the time dependent dynamics of the mean to include its estimation within the filtering framework and still use all available data.

In applying the storm counter method to the prediction problem it is sometimes necessary to estimate the mean and variance for storm counters with no sample and to estimate the storm counter itself at gages where it has not yet rained.

For example, suppose that a five-step prediction is being made for the three gages depicted in Figure 1 (five steps equal 1 hour). An estimate is needed for the mean and variance of rainfall at gages 4, 6, and 9 at 4:00 (1 hour later). The required storm counters are 20 for gage 4, 26 for gage 9, and 32 for gage 6. The sample sizes are 2 for storm counter 20, 1 for storm counter 26, and zero for storm counter 32.

To estimate the mean and variance for a storm counter with small (or zero) samples, data from adjacent storm counters is added until an acceptable sample size is reached. A lower limit on acceptable sample size must be set. If a storm counter  $n$  has a sample size smaller than the minimum allowed, the sample values for the next higher and next lower storm counters are included ( $n \pm 1$ ). The process is repeated ( $n + 2$ ,  $n + 3$ , ...) until the minimum sample size is achieved. A very large minimum sample size would force all storm counters to have the same mean and variance, i.e., a stationary mean and variance.

In order to predict the mean or variance using the storm counter method it is necessary to know the value of the storm

counter for the point being predicted. If it has begun to rain at the prediction point, the storm counter is simply the number of time steps since it started to rain. If it has not started to rain, it is necessary to estimate the time when it will begin to rain. In the storm counter framework, estimation of storm arrival is equivalent to estimation of the value of negative storm counters.

Several techniques were investigated to estimate storm arrival. They all have in common a fairly high uncertainty in the estimated arrival time. The methods tried fall into two broad categories: storm velocity methods and regression methods.

The storm velocity is estimated as part of the prediction scheme [see *Johnson and Bras*, 1978b]. The storm velocity defines an 'upwind' direction from each location where the storm arrival must be estimated. Looking along the upwind direction, the storm front might be found and its arrival time estimated. Estimating storm arrival with a velocity method requires a degree of spatial detail that was not available for the rain gage networks that form the test cases, and these networks are quite dense.

It is desirable to have a measure of uncertainty in the storm arrival prediction schemes. The chosen methodology follows.

Since it has rained at some locations, there will be some points with positive storm counters. These points can be used to estimate the parameters of an equation

$$L_i = a_0 + a_1x_i + a_2y_i + e_i \quad (23)$$

where

- $L_i$  storm counter at  $i$ th gage;
- $x_i, y_i$  location of  $i$ th gage;
- $e_i$  error at  $i$ th gage.

The estimated parameters can be used to estimate the storm counter at gages where it has not yet rained. Presumably, the predicted storm counter will not be positive at locations where it has not rained yet. It should be emphasized that the regression method requires an extrapolation of the function chosen, i.e., parameters will be estimated from gages with positive storm counters and extrapolated to locations with negative storm counters.

Naturally, (23) will describe storm arrival better for some types of events (for example, frontal storms) than for others (for example, convective cells). In essence, it is difficult to predict the arrival of a 'rough' or cell-type storm, and this difficulty will be reflected in the goodness of fit of (23). The model accounts for the uncertainty in the predicted storm arrival in the following way. The mean value (and variance) at a prediction point where it has not rained is found as a weighted sum of the mean values (variance) from a number of storm counters. The storm counters considered are centered about the estimate from (23) and extend two standard errors away from that estimate. The weights are from a normal density function centered at  $L_i$  from (23). Any positive storm counter values are ignored (the storm is known not to have arrived).

Occasionally, it may happen that all the values within two standard errors of (23) are positive; i.e., it should already be raining. In these cases the storm is considered to have 'missed' the prediction point; the storm counter is considered to be  $-\infty$ ; i.e., the storm will never arrive.

The effect of using a weighted mean and variance instead of a single estimate is to 'hedge' the uncertainty in storm arrival. When (23) produces a very good fit, the storm counters con-

sidered will fall in a narrow band about the estimate from (23). When (23) fits poorly, more storm counter values will be included in the weighted mean.

#### Estimation of the Covariance of Normalized Residuals

A complete discussion of the possible alternatives of covariance estimation in real time can be found in the work of *Johnson and Bras* [1978b]. For the examples presented in this paper a functional covariance form was assumed, requiring parameter estimation in real time.

It is reasonable to assume that for any time lag  $\tau$  the covariance of the normalized residuals will have a maximum value at some particular offsets ( $\Delta x_{\max}(\tau)$ ,  $\Delta y_{\max}(\tau)$ ). For lag zero,  $\Delta x_{\max} = \Delta y_{\max} = 0$ ; i.e., the covariance function has a maximum at the origin. If the storm is moving, it should be less variable in the coordinate system that moves with the storm than in a coordinate system fixed to the ground. This argues that relative maxima will all be in a straight line, i.e.,

$$\begin{aligned} \Delta x_{\max}(\tau) &= U_x \tau \\ \Delta y_{\max}(\tau) &= U_y \tau \end{aligned} \quad (24)$$

where  $U_x$  is the  $x$ -direction component of storm velocity and  $U_y$  is the  $y$ -direction component of storm velocity. There is evidence of this behavior in rainfall [*Marshall*, 1977; *Zawadzki*, 1973].

Equation (24) assumes that the storm velocity components are constant. If the storm velocity varies with time, the effect on the covariance is quite complex. A variable storm velocity does more than simply relocate the covariance maxima at each lag so that they are no longer collinear. A variable storm velocity implies that the location of covariance maxima are a function of absolute time, not just time lag. In short, variable storm velocity implies nonstationarity of the covariance, which would make its estimation much more difficult.

If the covariance decreases uniformly in all directions from the maxima defined by (24), it is isotropic in a coordinate system that moves with the storm. It is important to realize that isotropy in the moving coordinates implies anisotropy in a fixed coordinate system. The hope is that the storm's movement is the major source of anisotropy.

The distance from the maxima of (24) is measured by the storm distance  $d$ , defined

$$\begin{aligned} d(x_i, y_i, t_i; x_j, y_j, t_j) &= \{[(x_i - U_x t_i) - (x_j - U_x t_j)]^2 \\ &\quad + [(y_i - U_y t_i) - (y_j - U_y t_j)]^2\}^{1/2} \\ &= \{[\Delta x_{ij} - U_x \tau_{12}]^2 + [\Delta y_{ij} \\ &\quad - U_y \tau_{12}]^2\}^{1/2} = d_{ijr} \end{aligned} \quad (25)$$

In essence, the use of the storm distance reduces the dimensionality of the estimation. There are a variety of valid isotropic covariance functions [*Mejia and Rodriguez-Iturbe*, 1974; *Bras and Rodriguez-Iturbe*, 1976]. The covariance function used for this study is the exponential

$$E[\epsilon(x_i, y_i, t + \tau)\epsilon(x_j, y_j, t)] = \alpha_r e^{-\beta_r d_{ijr}} \quad (26)$$

Therefore at each prediction lead  $\tau$ , two parameters must be estimated,  $\alpha_r$  and  $\beta_r$ . Also required are estimates of  $U_x$  and  $U_y$ , the storm velocity components.

Clearly, the ideal approach would be to estimate all the covariance parameters simultaneously, at least,  $U_x$ ,  $U_y$ ,  $\alpha_0$ ,  $\beta_0$ ,

$\alpha$ , and  $\beta$ . In a multilead prediction format other height and decay parameters must be added to the list. The 6-parameter (or 8 or 10 or more) nonlinear estimation problem is formidable, especially when it must be solved in real time.

The problem can be greatly simplified if independent estimates of  $U_x$  and  $U_y$  are available, and this is the approach taken here. With  $U_x$  and  $U_y$  assumed known (velocity estimation will be discussed later), the covariance parameters  $\alpha$ , and  $\beta$ , can be estimated separately for each lag  $\tau$ . The following constraints must be imposed on  $\alpha$ , and  $\beta$ , to insure the validity of (26),  $\alpha \geq 0$ ,  $\alpha \leq \alpha_0$ , and  $\beta > 0$ .

The procedure used to estimate the covariance parameters  $\alpha$ , and  $\beta$ , is the following, define

$$\hat{\epsilon}_j = \frac{q_j - \hat{m}_j}{\hat{\sigma}_j} \quad (27)$$

where

$\hat{m}_j$  estimate of mean at gage  $j$  at time step  $t$  (from storm counter method);

$\hat{\sigma}_j$  estimate of standard deviation at gage  $j$  at time step  $t$  (from storm counter method);

$q_j$  rainfall rate (measured) at gage  $j$  at time step  $t$ ;

$\epsilon_j$  estimate of normalized residual at gage  $j$  at time step  $t$ .

If it has not rained at gage  $j$  at time step  $t$ , then  $\hat{m}_j$  is zero and  $\hat{\sigma}_j$  is undefined. The covariance estimation scheme must account for these undefined values.

Sample covariance values are estimated by the following equation:

$$\hat{c}_{ij\tau} = \frac{1}{T_1(i, j)} \sum_{t=T_2(i, j, \tau)}^{T_3(i, j, \tau)} \hat{\epsilon}_{i+t} \hat{\epsilon}_j \quad (28)$$

where

$\hat{c}_{ij\tau}$  estimates of covariance between gage  $i$  and gage  $j$  at lag  $\tau$ ;

$T_1(i, j)$  length of shortest record at gage  $i$  or  $j$ , i.e., maximum number of data values included in  $\hat{c}_{ij\tau}$  at any lag  $\tau$ ;

$T_2(i, j, \tau)$ ,  $T_3(i, j, \tau)$  start and end time steps so that all  $\hat{\epsilon}_{i+t}$  and  $\hat{\epsilon}_j$  in (28) exist.

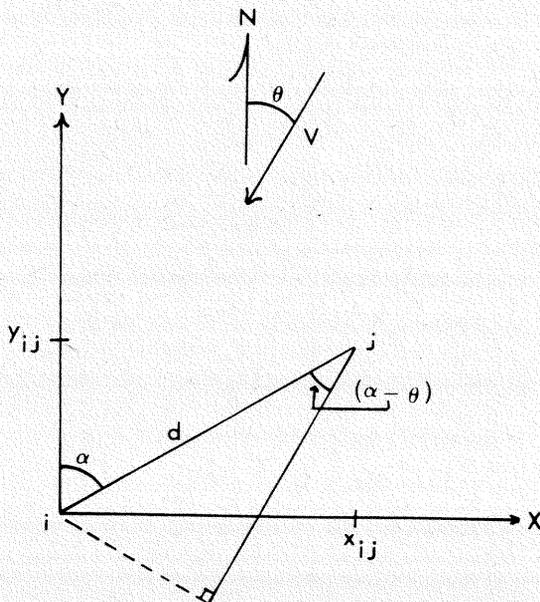


Fig. 2. Storm passing over gages  $i$  and  $j$  [from Marshall, 1977].

The quantity  $T_3 - T_2$  gives the number of data values included in a particular estimate  $\hat{c}_{ij\tau}$ , defined

$$n_{ij\tau} = T_3(i, j, \tau) - T_2(i, j, \tau) \quad (29)$$

Presumably, those values of  $\hat{c}_{ij\tau}$  corresponding to large  $n_{ij\tau}$  are more accurate on the average, so they should receive a higher weight in the procedure used to estimate  $\alpha$ , and  $\beta$ . A weighted sum of squares is formed:

$$\phi(\alpha, \beta) = \sum_{\substack{\text{all } t \\ \text{all } j \neq i}} n_{ij\tau} (\hat{c}_{ij\tau} - \alpha e^{-\beta d_{ij\tau}})^2 \quad (30)$$

The estimates of  $\alpha$ , and  $\beta$ , are then defined as the values which minimize (30),

$$\hat{\alpha}, \hat{\beta} = \min_{\alpha, \beta} \phi(\alpha, \beta) \quad (31)$$

The solution of (31) is a straightforward problem in nonlinear optimization solved via a Newton-Raphson technique. Having estimated the covariance parameters  $\hat{\alpha}$ , and  $\hat{\beta}$ , for  $\tau = 0, 1, \dots$ , it is a simple matter to produce the necessary covariance matrices  $D(t, t)$ ,  $D(t + \tau, t + \tau)$ , and  $D(t + \tau, t)$ . Of course, the storm distance cannot be computed unless the storm velocity is known.

#### Storm Velocity Estimation

It is difficult to estimate the storm velocity from point data [see Johnson and Bras, 1978b]. Trying to find the cross-covariance maxima leads to a high-order nonlinear parameter estimation problem. Rain gage data lack the spatial resolution to track storm patterns accurately. What is required is a linear estimation technique that relies on the temporal (not spatial) detail of rain gage records.

It is possible to estimate the storm velocity via a linear regression. The approach was first used by Marshall [1977] and has been applied by at least one other investigator [Shearman, 1977].

Referring to Figure 2, consider two rain gages,  $i$  and  $j$ , a distance  $d$  apart. A storm from direction  $\theta$  moving at speed  $V$  will take  $t_{ij}$  to travel between  $i$  and  $j$  in the direction of storm movement where

$$t_{ij} = \frac{d \cos(\alpha - \theta)}{V} \quad (32)$$

Expanding  $\cos(\alpha - \theta)$  gives

$$t_{ij} = \frac{\cos \theta}{V} \cdot d \cos \alpha + \frac{\sin \theta}{V} d \sin \alpha = \frac{\cos \theta}{V} \cdot y_{ij} + \frac{\sin \theta}{V} x_{ij} = b_1 y_{ij} + b_2 x_{ij} \quad (33)$$

where

$x_{ij} = (x_j - x_i)$  equals the  $x$ -direction distance from gage  $i$  to gage  $j$ ;

$y_{ij} = (y_j - y_i)$ ;

$V$  magnitude of storm velocity;

$\theta$  direction of storm movement (from north axis).

If  $t_{ij}$  can be estimated from the rain gage data, then the parameters of (33) can be estimated from linear regression. From the estimated parameters  $\hat{b}_1$  and  $\hat{b}_2$ , estimates of  $V$  and  $\theta$  are found

$$\hat{V} = (\hat{b}_1^2 + \hat{b}_2^2)^{-1/2} \tag{34}$$

$$\hat{\theta} = \tan^{-1} (\hat{b}_2/\hat{b}_1) \tag{35}$$

Finally, the two storm velocity components can be estimated as

$$\hat{U}_x = -\hat{V} \sin \hat{\theta} \tag{36}$$

$$\hat{U}_y = \hat{V} \cos \hat{\theta} \tag{37}$$

How should  $t_{ij}$  be estimated? The basic idea is to find the best match between the rainfall records at gages  $i$  and  $j$ . The time lag giving the best match is used to estimate  $t_{ij}$ . Marshall [1977] uses the cross-correlation function as a matching criterion. The authors use a slightly different matching criterion based on absolute differences, which offers computational advantages [Johnson and Bras, 1978b]. No matter what criterion is used, it will not always determine the true value of  $t_{ij}$ . Since the sample size tends to be large and the computational effort high, it is reasonable to avoid estimates which are likely to be in error. Several criteria were established to try to improve the quality of the sample used to estimate (33). Robust regression was used to handle the occurrence of outliers. The velocity estimation performed well in both synthetic and real storm data. The interested reader is referred to the work by Johnson and Bras [1978b] for a detailed discussion of the procedure and its behavior.

EXAMPLES AND RESULTS

Measures of Effectiveness

A variety of statistics can be computed which measure the accuracy or goodness of the prediction. Two commonly used statistics are the root mean square of the errors (rmse) and the coefficient of efficiency (CE). Then rmse is simply

$$\text{rmse} = \left( \frac{1}{T} \sum \xi_t^2 \right)^{1/2} \tag{38}$$

where  $\xi_t$  is the prediction error observed at time  $t$  and  $T$  is the number of time steps.

The coefficient of efficiency is a measure of the variance explained by the prediction and is computed via

$$\text{CE} = 1 - \frac{\sum \xi_t^2}{\text{Var}(i)} \tag{39}$$

where  $\text{Var}(i)$  is the variance of the rainfall at a given point. The statistic CE is similar to the multiple correlation coefficient of linear regression ( $R^2$ ). Both measure a percentage of

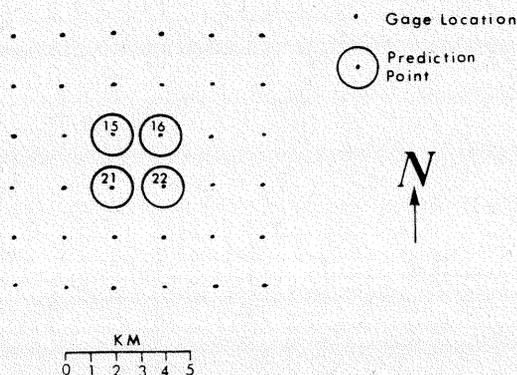


Fig. 3. Synthetic storm gage locations.

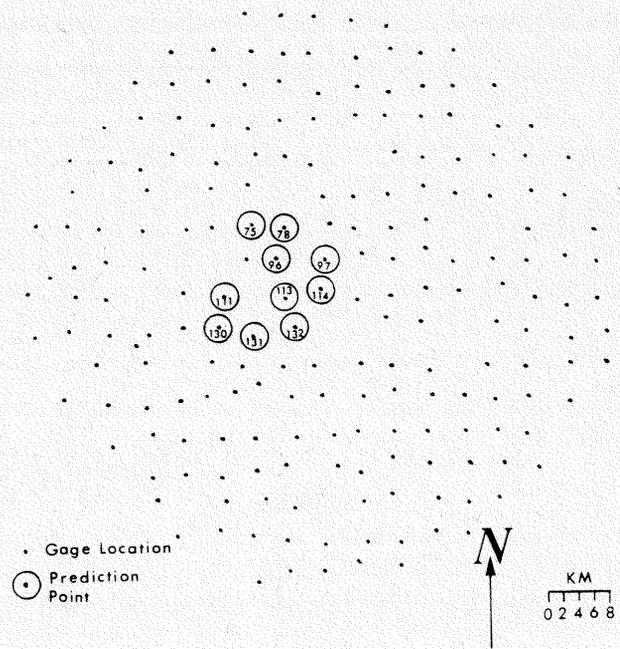


Fig. 4. Metromex gage locations.

variance explained. The  $R^2$  of linear regression always falls between 0.0 and 1.0. Likewise, CE has an upper bound of 1.0, but it does not have a lower bound; i.e., negative CE values are possible. In this context a negative value of CE indicates that the predictions have a higher variance than the actual rainfall. This occurs in some of the test cases. If it never rains, i.e.,  $r_t \equiv 0$  for all  $t$ , then CE is undefined.

The underlying problem with the CE and rmse statistics is that they treat errors at all time steps equally. This is perfectly

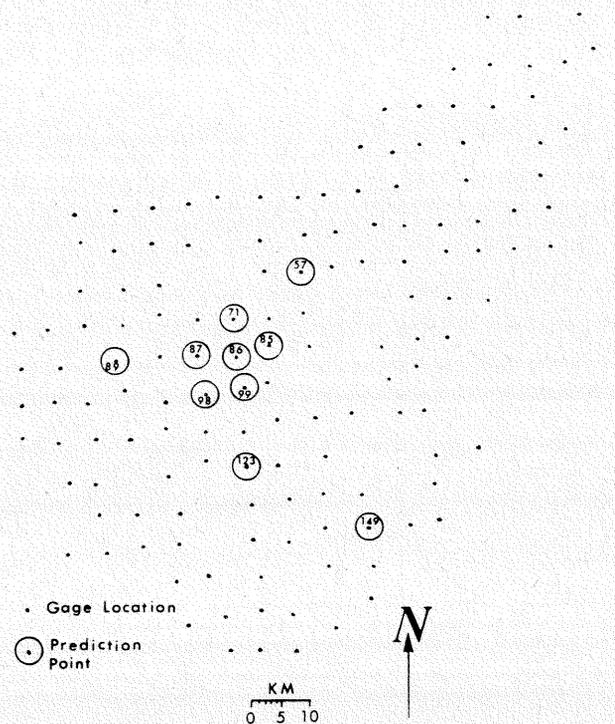


Fig. 5. Chickasha gage locations.

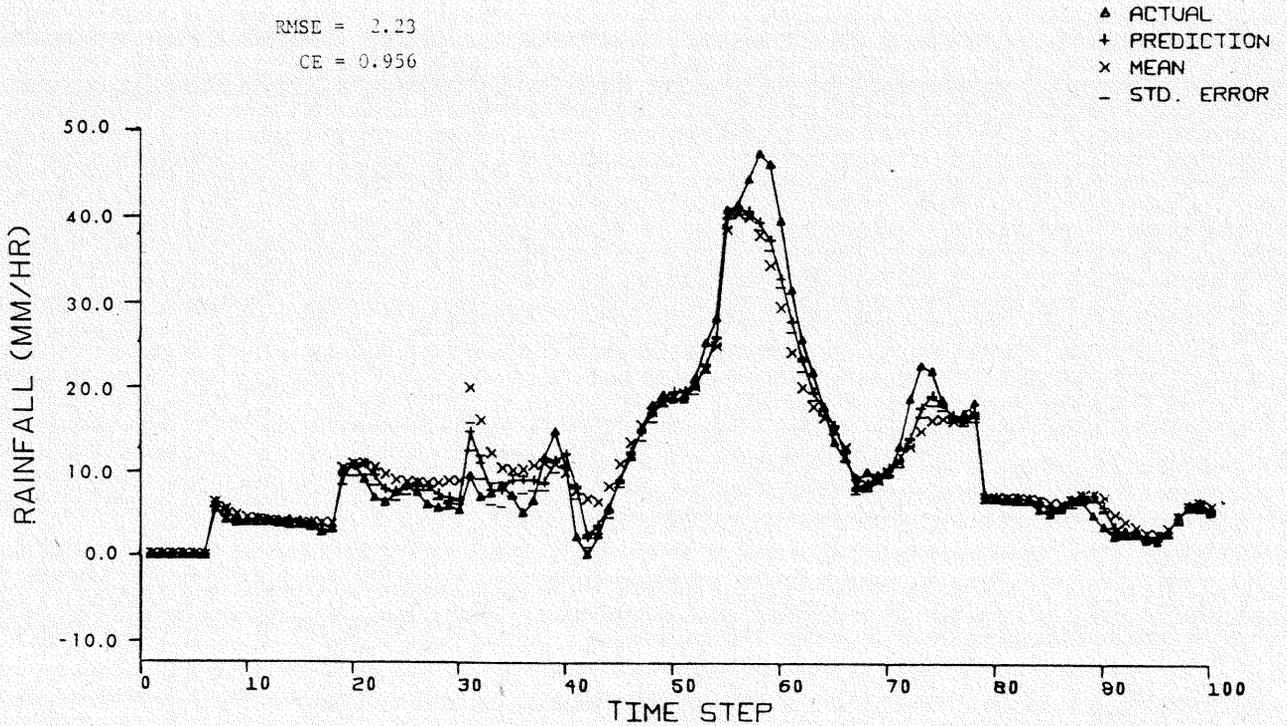


Fig. 6. Synthetic gage 16 lead 1.

reasonable for a stationary process, but rainfall is a transient process. The CE and rmse statistics are reported in the interest of allowing simple quantitative comparisons to be made. Rather than design more complicated statistics to measure the response of the prediction model, results will be presented in graphical form. There are four values shown on each plot: (1) the actual rainfall rate, (2) the predicted rainfall rate, (3) the predicted mean value of the rainfall rate as computed by the

storm counter method, and (4) the standard error of the prediction. The standard error comes from the prediction error covariance matrix,  $P(t + \tau|t)$  (equation (13)). When a single gage is considered, the standard error is simply the square root of the corresponding diagonal term of  $P(t + \tau|t)$ .

It is impractical to present a large number of results in a journal article. A more extensive discussion of test results for more storms can be found in the work of *Johnson and Bras*

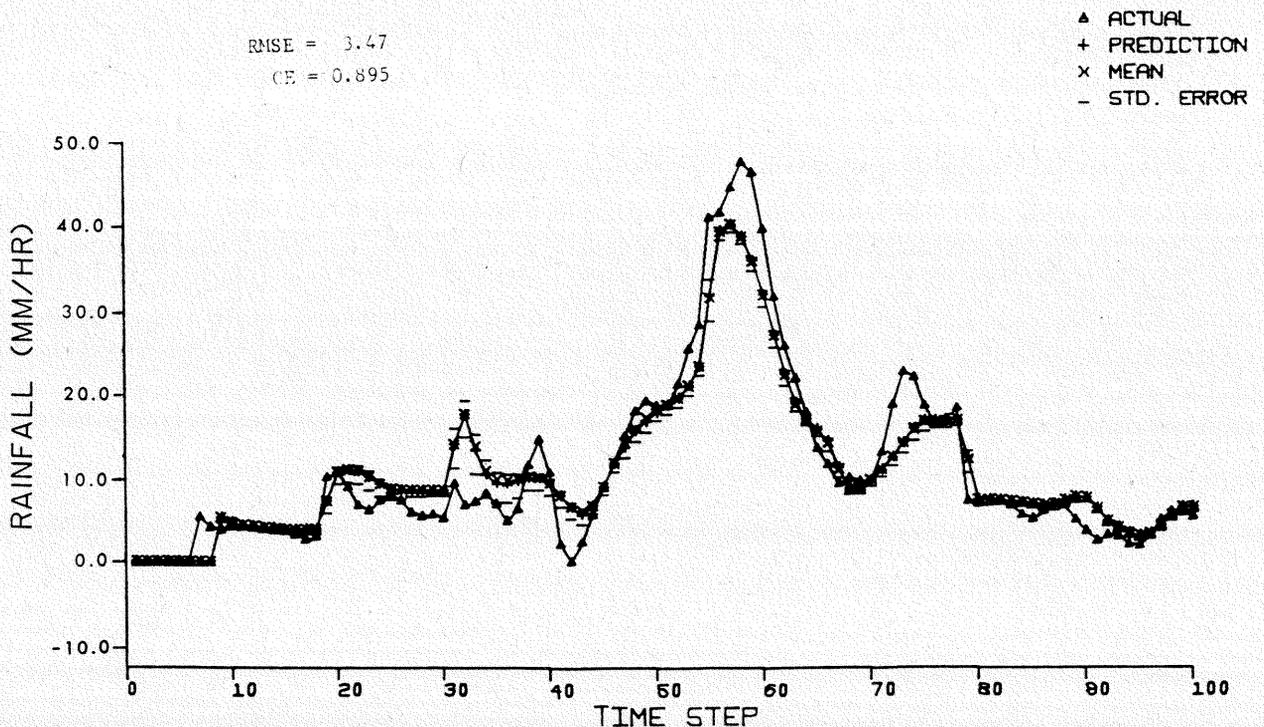


Fig. 7. Synthetic gage 16 lead 6.

RMSE = 8.63

CE = 0.510

▲ ACTUAL  
+ PREDICTION  
× MEAN  
- STD. ERROR

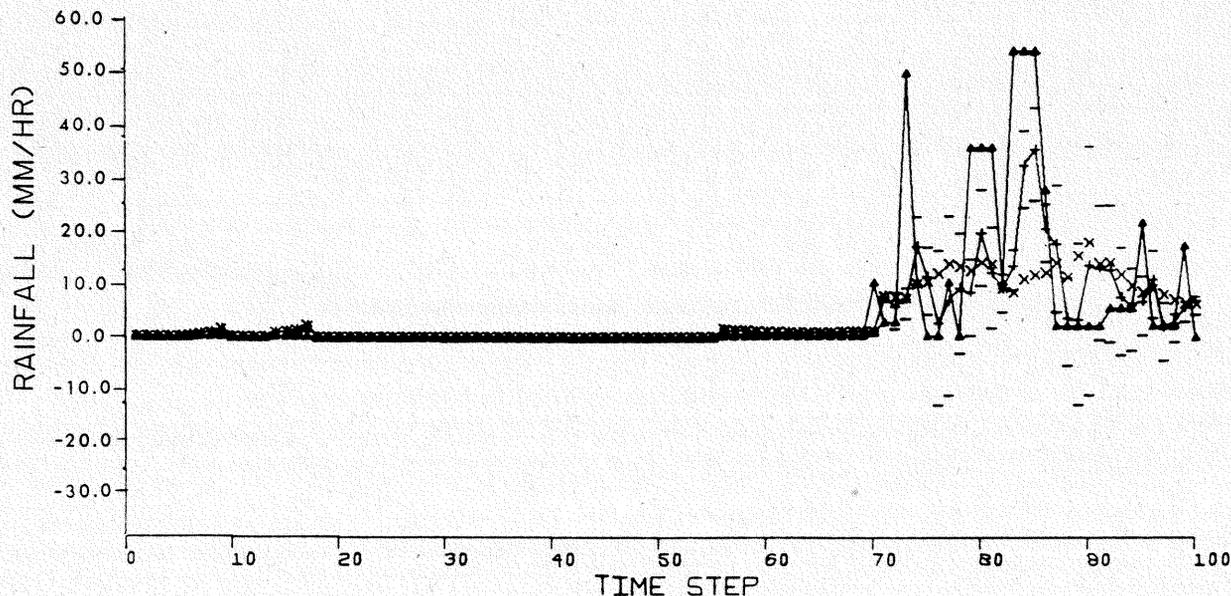


Fig. 8. CK060162 gage 87 lead 1.

[1978a]. The results presented here show single gages only, even though the multivariate nature of the model allows areal average predictions to be made. Because predictions are available at multiple leads, total volume predictions can also be made [see Johnson and Bras, 1978a].

Data Sources

The rainfall data used in the test cases below come from three sources: a rainfall synthesis model [Bras and Rodriguez-

Iturbe, 1976], the Southern Great Plains Watershed Research Center in Chickasha, Oklahoma, and Project Metromex, Illinois State Water Survey, Urbana, Illinois.

The synthetic event was sampled at 5-min time steps at the 36 rain gages shown in Figure 3.

The Metromex network is shown in Figure 4. There are 200 gages in this network with full records available. The gages are standard National Weather Service weighing bucket gages, nearly all with 24-hour charts (J. L. Vogel, personal

RMSE = 10.69

CE = 0.248

▲ ACTUAL  
+ PREDICTION  
× MEAN  
- STD. ERROR

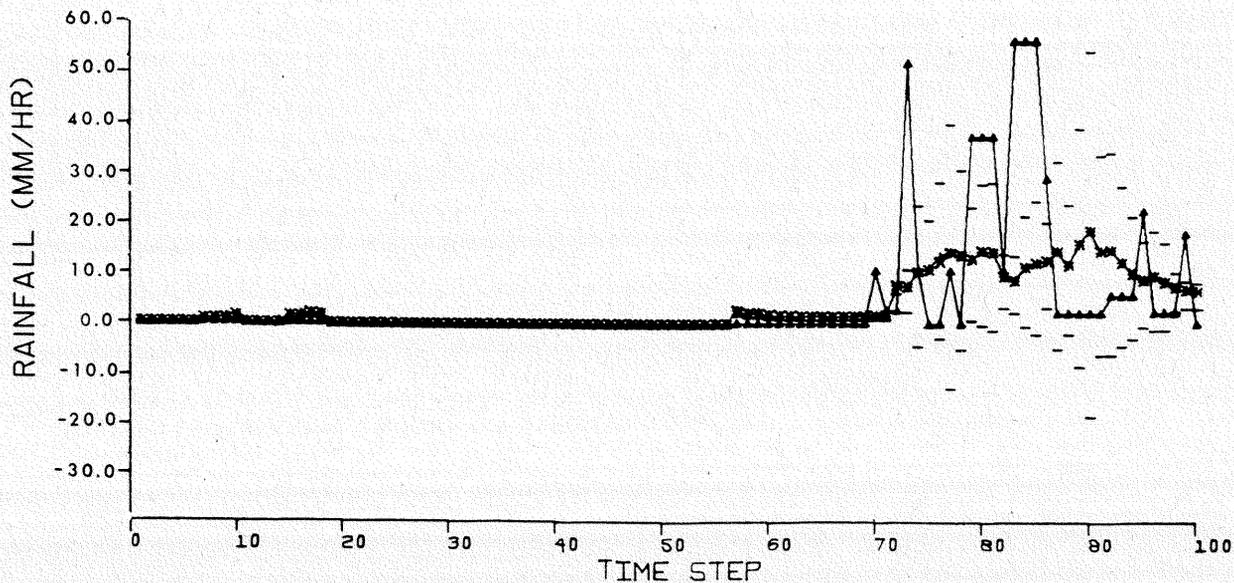


Fig. 9. CK060162 gage 87 lead 2.

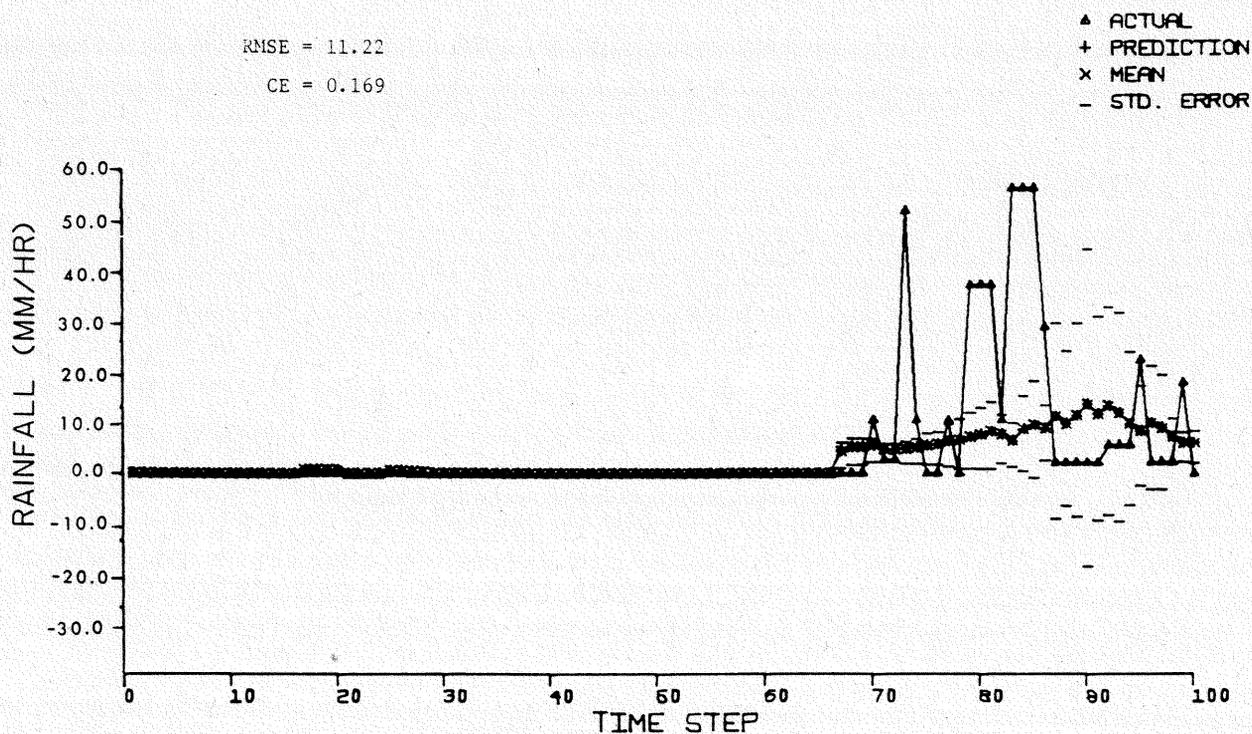


Fig. 10. CK060162 gage 87 lead 12.

communication, 1977). The data were already available in 5-min rainfall rates.

The Chickasha network is shown in Figure 5. There are 142 gages having full records for both test storms. The Chickasha network also uses weighing bucket gages. Data were obtained in breakpoint form and converted to 5-min rainfall rates.

*Synthetic Event*

The rainfall synthesis model [Bras and Rodriguez-Iturbe, 1976] produces events with precisely the characteristics assumed in the parameter estimation process. The mean and variance are nonstationary and are identical at each gage in

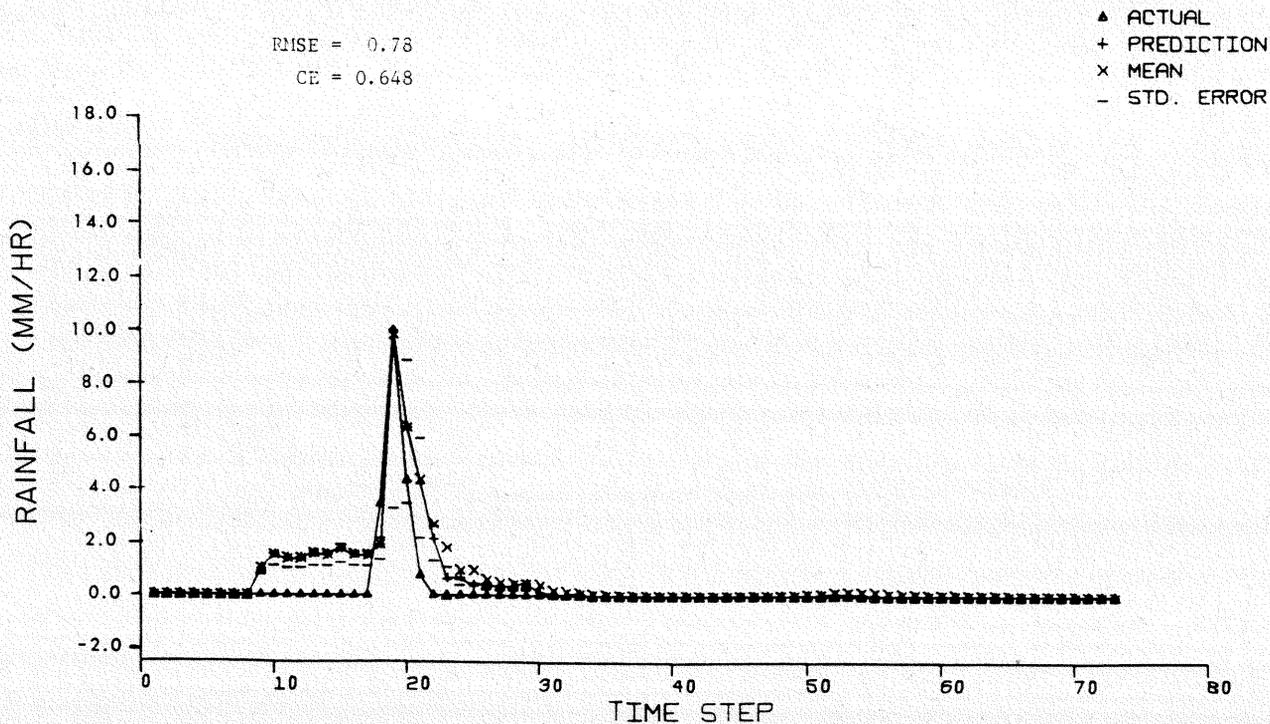


Fig. 11. MM081375 gage 111 lead 1.

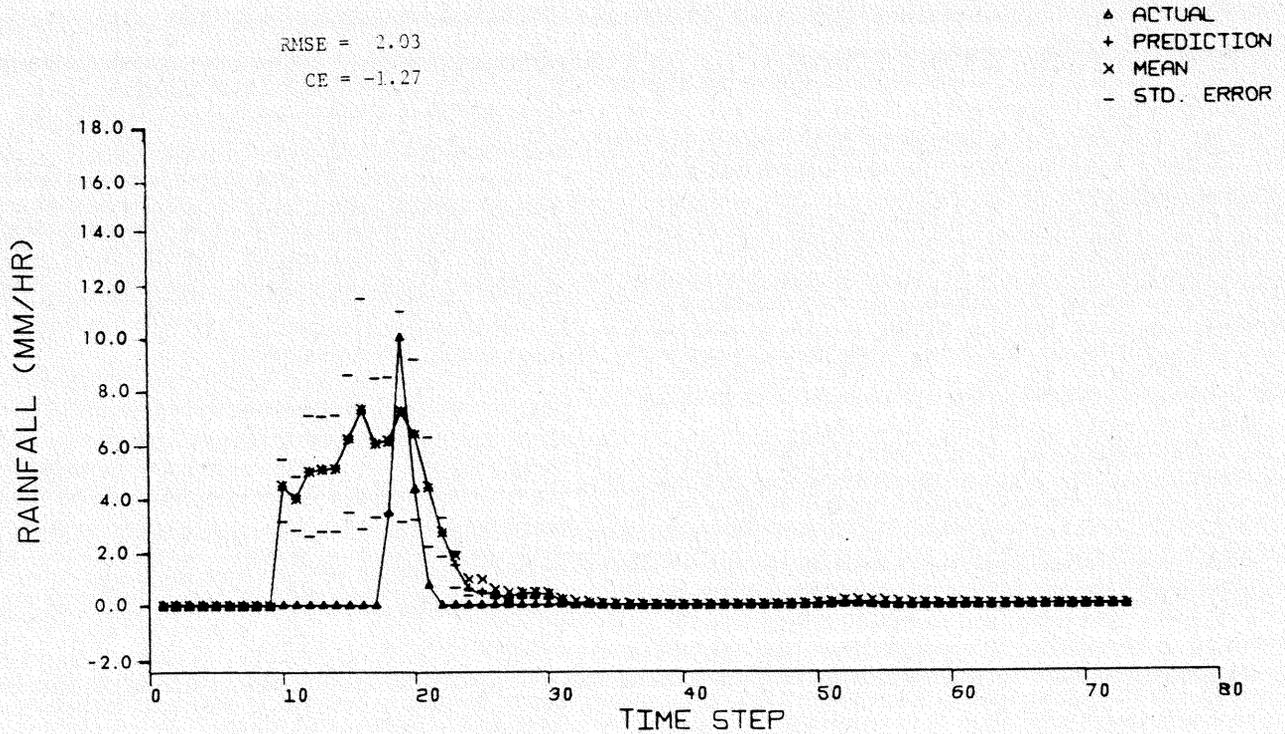


Fig. 12. MM081375 gage 111 lead 2.

relation to the time of storm arrival at that gage. The storm has a constant velocity with an isotropic exponential covariance in the moving frame of reference. Good performance for the synthetic event would indicate successful parameter estimation and will test expected model behavior.

The locations of the gages for the synthetic storm have been shown in Figure 3. Predictions were made at four gages (15,

16, 21, and 22) at lead values of 1, 2, 3, 4, 5, and 6 time steps (5, 10, 15, 20, 25, and 30 min). Figures 6 and 7 show the predictions for gage 16 at lead values of 1 and 6 time steps.

By almost any criterion the predictions are excellent. The accuracy of the prediction is degraded at higher leads but not severely. A number of other observations about model performance are important.

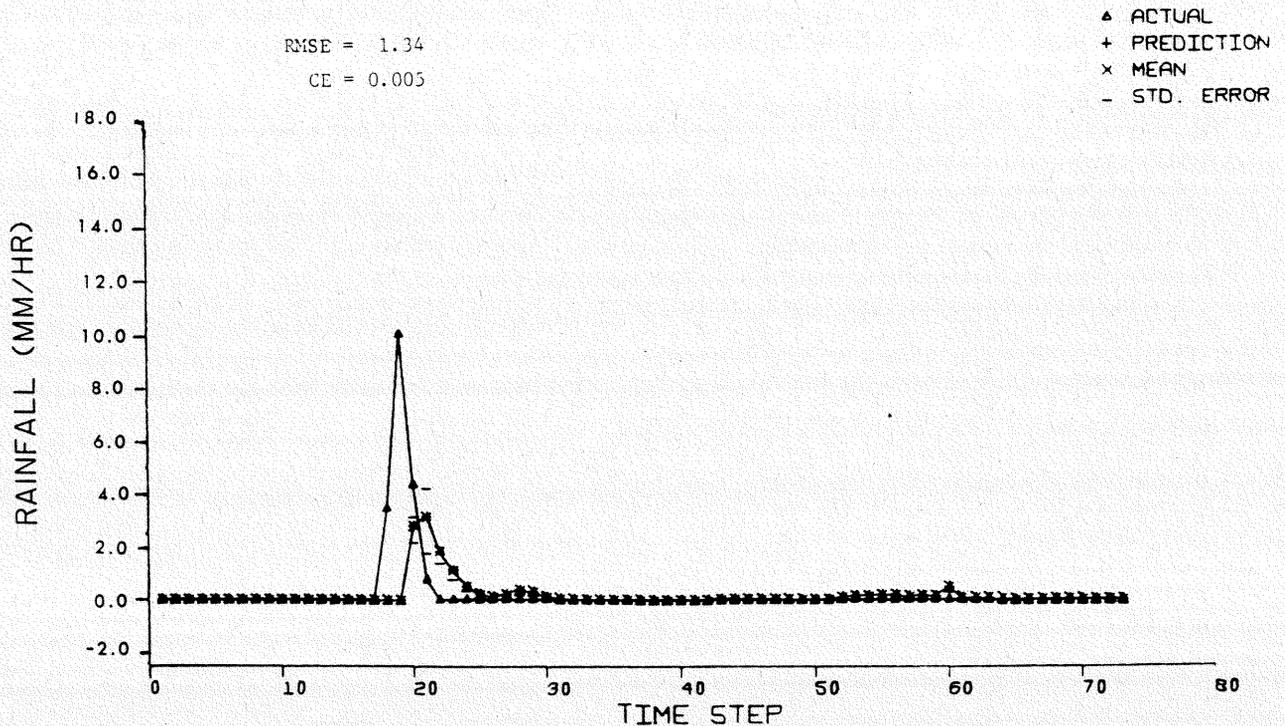


Fig. 13. MM081375 gage 111 lead 12.

1. The mean value identified by the storm counter method captures a lot of the structure of this event, reducing the burden on the residual forecasting technique.
2. The model does a fine job of predicting storm arrival for this case.
3. The model actually predicts rather than follows the rainfall, for example, notice the sharp drop in rainfall rate at time step 80.

#### *Storm CK060162*

Storm CK060162 occurred over the Chickasha network shown in Figure 5 on June 1, 1962. Predictions were made at 10 gages: 57, 71, 85, 86, 87, 89, 98, 99, 123, and 149. Predictions were made for leads of 1, 2, 3, 4, 5, 6, and 12 time steps (5, 10, 15, 20, 25, 30, and 60 min).

Figures 8–10 show predicted rainfall rates at gage 87 at leads of 1, 2, and 12 time steps, respectively. The rough and spiky nature of the hyetograph at gage 87 is characteristic of the hyetographs at other gages. As a result, the covariance decays rapidly, and higher lead values are not able to capture the finer structure of the hyetograph. Only the lead one prediction is able to capture the spikes at time steps 74, 79–81, and 83–85. However, the predictions at other leads do produce a reasonable 'smoothed' value of the measured rainfall.

#### *Storm MM081375*

The rain gage network for storm MM081375 is shown in Figure 4. Predictions were made at 10 gages (75, 78, 96, 111, 113, 114, 130, 131, and 132) for seven lead values (1, 2, 3, 4, 5, 6, and 12). The storm is described as a 'convective summer storm with variability in cell motion' (J. L. Vogel, personal communication, 1977), and such a storm should be difficult to predict.

The difficulty is exemplified by the rainfall records at gages 130 and 131 (approximately 3 mi apart). It rains for 20 min at gage 130, but it never rains at gage 131.

The linear model of storm arrival (equation (23)) fits storm MM081375 poorly. This forces the model to hedge against the possibility of storm arrival. At gages such as 131, where it never rains, the model does eventually decide that the storm has missed the gage, but not before several time steps with positive rainfall rates are predicted.

Considering the difficulty inherent in predicting this storm, the model performs well. Figures 11–13 show the response at increasing leads at a single gage where it does eventually rain. Notice the hedging behavior before storm arrival. Also notice the small value of the residual term; in most cases the prediction is equal to the predicted mean.

The lead 12 prediction at gage 111 (Figure 13) illustrates two characteristics of model behavior. The first time step with enough data to try to predict storm arrivals is time step 8. At that time the storm arrival fit is poor, so the model considers the possibility of storm arrival at time steps 9, 10, 11, ..., 20. It happens that the storm does arrive at step 20, which makes the prediction good for lead 12.

The second characteristic is more representative of general model behavior. Notice that it is raining at gage 111 at time step 20 (Figure 13). The predicted rainfall for time step 32 is practically zero. The point is that the model recognizes the short-lived nature of the convective event. Heavy rainfall rates in the present do not foreshadow heavy rainfall an hour in the future for this type of storm, and the model recognizes this.

## CONCLUSIONS

A rainfall prediction model has been developed which simultaneously predicts rainfall rates at multiple locations for multiple values of prediction lead. All model parameters are estimated solely from telemetered rain gage data for the event being predicted [see *Johnson and Bras*, 1978b].

The model includes velocity and direction of storm movement as explicit parameters. The storm arrival time at each predicted point is likewise an explicit parameter, which is estimated a priori for each location. The mean rainfall rate is not modeled as being either homogeneous spatially or stationary (constant with time). Likewise, the variance of rainfall is non-homogeneous and nonstationary.

The degree of prediction difficulty varies from one event to another. Naturally, the performance of the prediction scheme is dependent on the characteristics of the storm being predicted. The most critical issue is the ability to predict storm arrival. The method used to estimate storm arrival times works much better for some events than others; thus the prediction accuracy early in the event varies. When the storm arrivals fit the linear hypothesis used, the model is able to predict the start of rainfall with some degree of accuracy (for example, the synthetic event and storm CK060162). When storm arrivals do not fit a linear form well, the model hedges the uncertainty in storm arrival and generally produces a reasonable, although less accurate, prediction.

As expected, the accuracy of the prediction generally decreases with increasing prediction lead. Except for very short prediction leads (for example, 5–10 min) it is impossible to capture the fine structure of rainfall variability. At high leads the covariance decay is rapid enough that the predicted residual value is usually zero. As a result, the nonstationary mean determines the structure of the forecast for higher leads. Therefore the more the estimated mean value describes the structure of the event, the better the forecast. The model is able to recognize the short-lived nature of a convective event (for example, MM081375) and yet is able to produce a reasonable smoothed out version of other types of events (for example, CK060162) for higher lead prediction.

The ability of the model to predict the start of the storm and the end of the storm is a direct result of the decision to use a nonstationary model for storm behavior. The accuracy of the prediction generally increases with real time. Assuming that it rains at all predicted points, there is eventually no uncertainty regarding the storm arrival time. As the storm proceeds, more data are collected, allowing better parameter estimates.

Early in the event, little data is available, and default values must be assumed for some parameters—particularly, the covariance parameters. In the current study these default values were set rather arbitrarily, but experience with model application would allow more appropriate choice of default values and thus better behavior early in the event. Setting defaults is particularly important for smaller networks, which provide less storm data at each time step.

On the whole, the model performs well for the test cases considered and appears to be computationally feasible for a rain gage network of reasonable size.

*Acknowledgments.* The work was sponsored by the National Science Foundation under grant ENG 76-118-17 to the Massachusetts Institute of Technology, Department of Civil Engineering, Ralph M. Parsons Laboratory for Water Resources and Hydrodynamics. Help

in publication was provided by the National Weather Service, Hydrologic Research Laboratories.

## REFERENCES

- Bras, R. L., and I. Rodriguez-Iturbe, Rainfall generation: A non-stationary time-varying multidimensional model. *Water Resour. Res.*, 12(3), 450-456, 1976.
- Gelb, A. (Ed.), *Applied Optimal Estimation*, MIT Press, Cambridge, Mass., 1974.
- Grigg, N. S., J. W. Labadie, and H. Wenzel, Metropolitan water intelligence systems completion report, phase III. *Rep. NTIS PB 234432*, Colo. State Uni., Fort Collins, June 1974.
- Jamieson, D. G., and J. C. Wilkinson, River Dee research program. 3. A short-term control strategy for multipurpose reservoir systems. *Water Resour. Res.*, 8(4), 911-920, 1972.
- Johnson, E. R., and R. L. Bras, Short term rainfall prediction: A non-stationary multivariate stochastic model. *Tech. Rep. 233*, Ralph M. Parsons Lab. for Water Resour. and Hydrodyn., Dep. of Civil Eng., Mass. Inst. of Technol., Cambridge, Mass., April 1978a.
- Johnson, E. R., and R. L. Bras, Real time estimation of velocity and covariance structure of rainfall events using telemetered rain gage data—A comparison of methods. *J. Hydrol.*, in press, 1978b.
- Labadie, J. W., N. S. Grigg, and B. H. Bradford, Automatic control of large-scale combined sewer systems. *J. Environ. Eng. Div. Amer. Soc. Civil Eng.*, 101(EI), Feb. 1975.
- Marshall, R. J., Statistical analyses of storm and daily rainfall data. Ph.D. dissertation, Univ. of Bristol, Bristol, England, April 1977.
- Mejia, J. M., and I. Rodriguez-Iturbe, On the synthesis of random field sampling from the spectrum: An application to the generation of hydrologic spatial processes. *Water Resour. Res.*, 10(4), 705-711, 1974.
- Sage, A. P., and J. L. Melsa, *Estimation Theory With Applications to Communication and Control*, McGraw-Hill, New York, 1971.
- Shearman, J. R., The speed and direction of movement of storm rainfall patterns with reference to urban storm sewer design. *Hydrol. Sci. Bull.*, 22(3), Sept. 1977.
- Zawadski, I. I., Statistical properties of precipitation patterns. *J. Appl. Meteorol.* 12, 459-472, 1973.

(Received October 26, 1978);  
revised August 23, 1979;  
accepted September 10, 1979.)

