AGENCY FOR INTERNATIONAL DEVELOPMENT
WASHINGTON, D. C. 20523

BIBLIOGRAPHIC INPUT SHEET

FOR AID USE ONLY
0153

1. SUBJECT CLASSIFICATION

A. PRIMARY: Agriculture — AP10-0000-0000

B. SECONDARY: Water resources and management

2. TITLE AND SUBTITLE

Space-time validation of a thunderstorm rainfall model

3. AUTHOR(S)

Fogel,M.M.; Duckstein,Lucien; Kisiel,C.C.

4. DOCUMENT DATE	5. NUMBER OF PAGES	6. ARC NUMBER
1971	8p.	ARC

7. REFERENCE ORGANIZATION NAME AND ADDRESS

Ariz.

8. SUPPLEMENTARY NOTES (Sponsoring Organization, Publishers, Availability)

(In Water resources bul.,v.7,no.2,p.309-316)

9. ABSTRACT

10. CONTROL NUMBER

PN-RAA-097

11. PRICE OF DOCUMENT

12. DESCRIPTORS

Mathematical models
Periodic variations
Thunderstorms

13. PROJECT NUMBER

14. CONTRACT NUMBER

CSD-2457 211(d)

15. TYPE OF DOCUMENT

AID 590-1 (4-74)

Volume 7 - Number 2 - April 1971

6

Reprinted from the

Water Resources Bulletin

Journal
of the American Water Resources Association

Urbana, Illinois 61801
U.S.A.

SPACE-TIME VALIDATION OF A THUNDERSTORM RAINFALL MODEL[1]

M. M. Fogel, L. Duckstein and C. C. Kisiel[2]

ABSTRACT. A probability model for predicting the occurrence and magnitude of thunderstorm rainfall developed in the southwestern United States was tested in the metropolitan Chicago area with reasonable success, especially for the moderate to the extreme runoff-producing events. The model requires the estimation of two parameters, the mean number of events per year and the conditional probability of rain given that an event has occurred. To tie in the data from more than one gage in an area, an event can be defined in several ways, such as the areal mean rainfall exceeding 0.50 inch and at least one gage receiving more than 1.0 inch. This type of definition allows both of the model parameters to be obtained from daily warm-season rainfall records. Regardless of the definition used a Poisson distribution adequately described the number of events per season. A negative binomial distribution was derived as representing the frequency density function for rainfall where several gages are employed in defining a storm. Chicago data fit both distributions very well at events with relatively high return periods. The results indicate the possibility of using the model on a regional basis where limited amount of data may be used to estimate parameters for extensive areas.
(KEY WORDS: thunderstorm rainfall model; frequency analysis; probability distributions; parameter estimation)

INTRODUCTION

The effective management of urban water resources is contingent on rainfall-runoff-quality relationships and on a set of inputs whose time-spatial distributions are adequately represented. One such major input is the short-duration, high-intensity thunderstorm rainfall that produces many of the extreme warm season floods. Knowledge of this type of storm is also required for use with watershed models to determine the effects of various land treatments designed to alter the normal water cycle.

In a previous study, a model for determining point rainfall frequencies in convective-type storms was developed using 12 years of record from a dense rain gage network [Fogel and Duckstein, 1969]. The model was compared with long-term records of a nearby Weather Bureau station in Tucson, Arizona and was found to give good results. Essential parameters of the model are the mean number of storms events within a given area per unit of time and the conditional probability of rain at a point given that a storm has occurred in the area. This paper, as one of its goals, investigates the extension of the model to other geographic areas. A restraint placed on the model is that it must be able to use data that is readily available.

A second aim of this study concerns the length of record required to estimate the parameters of the model. The length of precipitation record deemed necessary to obtain a stable frequency distribution has been estimated as ranging from 30 to 50 years for most regions [Wiesner, 1970].

[1]Paper No. 71029 of the *Water Resources Bulletin* (Journal of the American Water Resources Association). Discussions are open until six months from date of publication.
[2]Professors of Watershed Management, Systems Engineering and Hydrology and Water Resources, respectively, University of Arizona, Tucson, Arizona.

Since thunderstorm rainfall is recognized to be more variable in time and space than other storm types, it seems likely that 50 years of data may be required for satisfactorily estimating the return period of a given amount of rain. This investigation, therefore, is concerned with the trade-off between the current use of an approximate model which may require only a few years of data for calibration and the future use of a possibly more accurate model.

The question may also be asked as to the value of an additional piece of information. Thus, parameter estimation is involved in making statistical inferences as well as in decision theory.

In summary, this paper sets forth a model for the prediction of high-intensity summer rainfall which may become regional in scope. If additional investigation substantiates the assumptions made in developing the model, the estimation of two regional parameters would permit the forecasting of extreme events within that region.

THUNDERSTORM RAINFALL MODEL

The model was developed with the hypothesis that the occurrence of a storm within a given area is independent of the maximum amount of rainfall that falls during a storm within this area. In addition it is assumed that each storm has a Gaussian spatial distribution of rainfall. The historical record of thunderstorm rainfall, therefore, can be considered as the result of many storms whose center depth, or point of maximum rainfall, is a random variable as is the distance between the storm center and the raingage in question. In other words, there is a probability associated with a center over a given area which is independent of the probability that the center will have a specified amount of rain. Summing up this joint probability in space produces the conditional probability of receiving a certain amount of rain given that a storm has occurred.

Analysis of the data from a rain gage network in the Tucson area indicated that an exponential model could explain 75 percent of the variance in the spatial distribution of rainfall. The data also tended to substantiate the assumptions by McDonald [1959] and Osborn and Reynolds [1963] that the locations of air mass thunderstorms in fairly level country may be regarded as randomly distributed. Using a uniform distribution of storm center locations and the exponential spatial distribution model in a deterministic manner, the conditional probability of point rainfall was found to be in close agreement with a geometric distribution. The parameter of this frequency distribution, p, represents the probability that given a storm in the area, the point in question within that area will receive a measurable amount of rain. Mathematically expressed, the frequency density function for x units of rain is

$$f_R(x) = (1-p)\,P^x \qquad x = 0, 1, \ldots \tag{1}$$

Although 12 years of record are not sufficient to establish the use of a distribution for an annual event, the data indicated a trend towards a Poisson-distributed random variable for the number of storm events per season. A similar result was obtained in a study by Todorovic and Yevjevich [1969] who have concluded that the number of storms in an interval is Poisson-distributed. The frequency distribution for j storms per year is therefore

$$f_N(j) = \frac{e^{-m}\, m^j}{m!} \qquad j = 0, 1, \ldots \tag{2}$$

where the parameter m is the mean annual number of storms.

The cumulative maximal distribution function for point rainfall, derived from the above two distributions, is

$$F_Q(x) = \sum_j \left\{ [F_R(x)]^j f_N(j) \right\} \tag{3}$$

where $F_R(x)$ is the cumulative conditional probability of obtaining less than x units of rain.

EXTRAPOLATION TO OTHER AREAS

To test the model in another geographic area, the rain gage network of metropolitan Chicago was selected since data for several gages are readily available for an extended period of time. While the model generates point rainfall frequencies, data from a network can be used to obtain the necessary parameters for the model. To do so required that a storm be defined in such a way that all gages are used. In a study of heavy warm-season rainfalls over a 400 square-mile network in Illinois, Huff [1967] defined a storm as one in which the areal mean rainfall exceeded 0.50 inch and/or one or more gages recorded over 1.0 inch. Huff in a later paper [1968] used a slightly different definition in which at least one gage had to receive more than an amount equal to the 2-year recurrence interval in addition to the requirement that the mean exceed 0.50 inch.

Five rain gages located within approximately a 50 square-mile area in metropolitan Chicago with a historical record of 19 years were selected for the test. In an attempt to simulate the occurrence of a storm as defined in the original model, several definitions were investigated as designating the occurrence of an event. These were as follows:

1. Network mean greater than 0.50 inch
2. Network mean greater than 0.50 inch and at least one gage more than 1.0 inch
3. Similar to (1) except 0.75 inch used
4. Similar to (2) except 0.75 inch and 1.25 inch used.
5. Any gage greater than 0.50 inch and the difference between any two more than 0.50 inch

As a means for determining the extent to which an arbitrary definition of the occurrence of an event is related to that used in the original model, Weather Bureau data in the immediate vicinity of the experimental rain gage network were examined. Upon consideration of four Weather Bureau stations in the Tucson area, an event was defined as one in which the mean areal rainfall exceeded 0.25 inches and at least one gage recorded more than 0.50 inch. With this definition, the assumption is made that the occurrence of an event is similar to that which was used in the development of the original model. The model stated that an event occurred when there was at least one storm center (point of maximum rainfall) within the rain gage network area. The arbitrary definition resulted in values for the parameters m and p of 6.0 and 0.46, respectively. This compares with 5.3 and 0.48 found in the model's original calibration period.

For the Chicago data, the mean number of yearly occurrences were determined in all cases. As in the model, the number of storms in an interval was fit to a Poisson distribution. Using the Kolmogoroff-Smirnov test for comparing the observed with the theoretical distribution it was found that the Poisson distribution could not be rejected for any of the cases at significance levels of at least 10 percent. Figure 1 compares the observed with the Poisson distribution for the second case, which is essentially the definition of a storm that Huff used. For illustrative clarity the theoretical distribution is shown as a continuous function while in fact the Poisson distribution is a discrete function.

In the model, a geometric distribution was found to adequately describe the conditional probability of point rainfall given that a storm has occurred. For a network of gages the

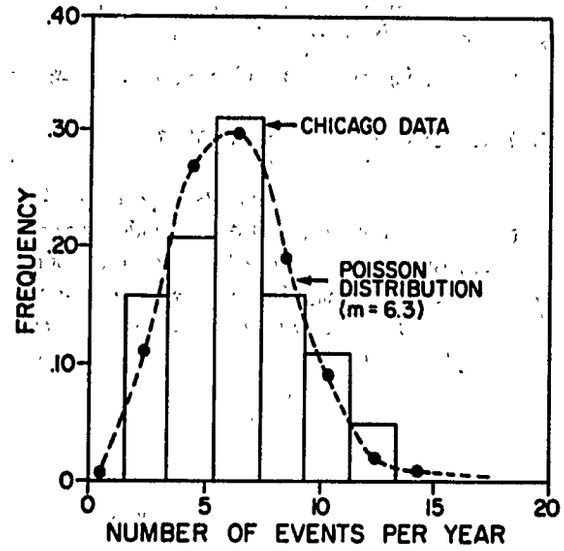

Figure 1. Distribution of annual number of events where areal mean rainfall exceeded 0.50 inch and at least one gage recorded more than 1.0 inch.

frequency distribution may possibly follow another function. For example the generating function for the geometric distribution given in Eq. (1) is

$$G(s) = \frac{1-p}{1-ps} \tag{4}$$

Assuming for the moment that there are r mutually independent gages in a network, the problem is to find the probability density function for the mean rainfall or alternatively, that of the sum of the rainfall amounts at all the gages. Since the gages are assumed to be independent the sum is a random variable composed of r random variables. The generating function for this sum according to Feller [1950] is

$$[G(s)]^r = \left(\frac{1-p}{1-ps}\right)^r \tag{5}$$

which is the generating function of the negative binomial distribution. In other words, the negative binomial distribution is the r-fold convolution of the geometric distribution and is expressed as

$$f(x) = \binom{r+x-1}{x} p^x (1-p)^r \qquad x = 0, 1, \ldots \tag{6}$$

where x now represents the mean rainfall of r independent gages and p is the same as in the geometric distribution. When the daily warm-season rainfall amounts recorded by a network such as in the Chicago metropolitan area are not considered as completely independent variables, r must assume some value between one and the number of rain gages.

To obtain p the Chicago records for one of the five stations were re-examined and the empirical probabilities of given amounts of rain were determined. These were plotted on semi-logarithmic paper and the line of best fit drawn (Figure 2). This line represents the theoretical

geometric distribution for given amounts of rainfall. The parameter p is determined from the complement of the vertical-axis intercept as this value represents the probability of non-occurrence of an event as defined above.

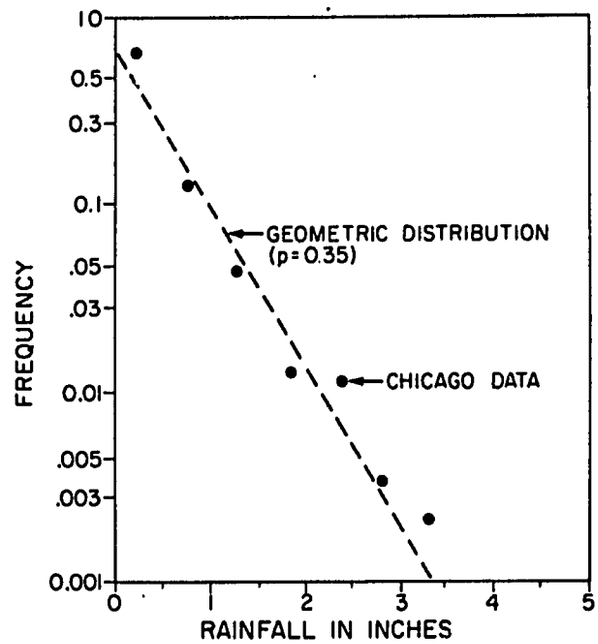

Figure 2. Distribution of point rainfall.

Using the definition for a storm as one in which the areal mean rainfall is greater than 0.50 inch and at least one gage records more than 1.0 inch, the empirical probabilities for selected rainfall amounts were determined. With the parameter p determined from Figure 2 and r set equal to 5, the number of gages, the negative binomial distribution was used to calculate the theoretical probabilities for given rainfall amounts. The comparison between the observed and the theoretical frequency distributions is shown in Figure 3.

While the theoretical and observed distributions have similar shapes, the agreement at the lower values is not too good. At the tail of the distribution, however, or where the probabilities of extreme events are determined, there is close agreement between the two distributions. It is this end of the distribution that has important implications in the design of water control structures.

An explanation of the difference between the theoretical and observed distributions can be found in the original assumption of mutual independence between the rain gages. If there is some dependence within the rain gage network, a value for r of three or four may be more realistic. Another possible explanation concerns the arbitrary definition of a storm which results in distributions that do not include all possible events.

PARAMETER ESTIMATION

An important consideration in the development of regional models concerns the period of

record necessary to estimate the parameter of a model. To study this aspect, the historical data of the experimental rain gage network in Tucson, Arizona were lumped together into three-year periods. Each period was analyzed individually to obtain the model parameters m and p. A three-year period was selected because a valid frequency analysis necessitates a minimum number of events per time interval. By taking the mean of three years, a normal distribution for that mean could be assumed even if the yearly distribution was not symmetric. As mentioned earlier, the distribution for the annual number of events was assumed to be Poisson. The independence of events underlying the Poisson distribution makes the parameter m an arithmetic mean of the annual values. The methodology for obtaining the parameters was the same for the three-year periods as was used in the development of the original model. Table 1 presents the range of values obtained for the two parameters.

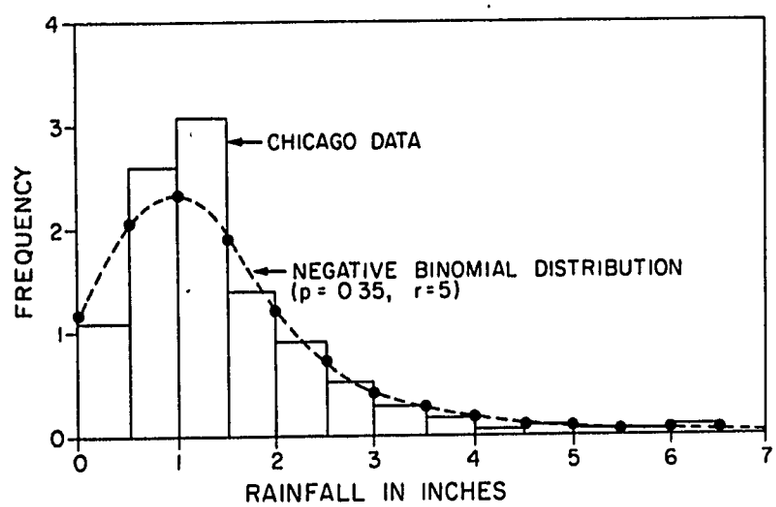

Figure 3. Distribution of areal mean rainfall.

TABLE 1. Values of Model Parameters for Three-Year Periods

Period	Parameter: Mean number of events, m	Parameter: Conditional probability of rain, p
1956-58	4.33	0.46
1959-61	7.33	0.50
1962-64	4.00	0.54
1965-67	5.67	0.18
1968-70	3.00	0.39

The question arises then, how many three-year periods are needed to have a "good" estimation $\bar{X}_n$ of the mean. Since it will depend on the use made of the mean two cases can be distinguished:

(1) a decision will be made on the basis of the estimated $\bar{X}_n$ of the mean (as well as of p); for example, the size of a flood protection structure will depend on a maximal value of $\bar{X}_n$ such as $\bar{X}_n + C_{95}$ where C_{95} is the 95 percent confidence interval limit; or else, the size of a reservoir will be chosen to yield enough water even if the mean takes on a minimal value such as $\bar{X}_n - C_{95}$. Clearly, the value taken on by the parameter p, the probability of point rainfall conditioned on the presence of a convective storm will also enter into the decision.

(2) an inference on $\bar{X}_n$ is necessary for scientific purposes; for example, to check a meteorological model describing the occurrence of convective storm cells.

The calculations shown in Table 2 may be used to help answer the above questions in either case. Confidence intervals for the means, $\bar{X}_n$, were calculated with the assumption that $\bar{X}_n$ is a normal variate.

TABLE 2. Confidence Interval About Mean Value of Parameter m for Various Periods of Record

Three-year periods, n	m	$\bar{X}_n$	Standard error of mean $\bar{X}_n$	95% Confidence Interval
1	4.33	4.33	-	-
2	7.33	5.83	1.50	± 9.46
3	4.00	5.22	1.06	± 3.10
4	5.67	5.33	0.77	± 1.80
5	3.00	4.87	0.75	± 1.62

In the first case, if a decision is to be taken, the questions which come up at a given point in time are:

- Can the decision be taken at all, knowing for example that $\bar{X}_n = 5.22 \pm 3.10$ with an a priori chance of 0.95 (n = 3)?
- Assuming the decision can be taken, how much is a reduction of uncertainty worth and what is the cost of waiting to obtain this reduction in uncertainty?

Note that at the point, n = 3, the extent of the uncertainty reduction is unknown if it is decided to wait until n = 4. This uncertainty reduction can be estimated either by using a regional approach (data obtained from another location similar to the one under study) or by a gross extrapolation, assuming for example that $\bar{X}_4 = \bar{X}_3$, or by using a Bayesian decision theory approach [Davis and Dvoranchik, 1970].

For the second case where a statistical inference is to be taken, a scientist desires to estimate a given physical parameter although there is no limit to the precision. Care must be taken to judge the worth of better precision on m when used in a model that also requires an inference on another parameter p which may be less accurately determinable than m A model that takes into account not only the number of events per year but also their magnitude variation is presently being developed.

In any case, it is most important to examine the sensitivity of either a decision or a model to the parameters. For the maximal and minimal distributions of point rainfall, the model was found to be more sensitive to the value of p than that of m; thus, as these distributions are fundamental for design purposes, more resources should be allocated to studying p than m. It is hoped that these brief considerations have demonstrated how the parameter estimation problem can be treated in a goal-oriented framework to provide more rational answers.

CONCLUSIONS

1. The annual number of runoff-producing rainfall events is a Poisson variate, irrespective of how an event is defined. Being a one-parameter distribution, the only information needed is the mean number of events per season.
2. The conditional probability of thunderstorm rainfall at a point is defined by a geometric distribution whose parameter is determined by the probability of having rain of at least a measurable amount given the occurrence of an event.
3. When a storm event is arbitrarily defined to involve rainfall at several points the frequency density function of the amount of rainfall follows a negative binomial distribution. This two-parameter distribution requires knowledge of the conditional probability of having rain and the number of data collection points or rain gages involved.
4. The use of a probability model for predicting the amount of thunderstorm rainfall on a regional basis appears to have some possibility. Frequency distributions concerning the number and magnitude of rainfall events developed in a thunderstorm model under southwestern United States conditions were used on metropolitan Chicago data with satisfactory results. Only daily warm season rainfall records were needed.
5. Parameters can be estimated for two purposes, i.e., to make an engineering decision or to draw a statistical inference. In either case, the value of obtaining a better estimate gives a stopping rule for taking data. That value is relative to economic considerations for the decision case and relative to the precision of other model parameters for the inference case.

ACKNOWLEDGEMENTS

The work upon which this paper is based was funded in part by the United States Department of the Interior, Office of Water Resources Research, as authorized under the Water Resources Act of 1964.

LITERATURE CITED

Davis, D. R. and W. M. Dvoranchik. 1970. Evaluation of the worth of additional data. A paper presented to the Sixth American Water Resources Conference, Las Vegas, Nevada, October 26-30.

Feller, William, 1950. An introduction to probability theory and its applications. John Wiley and Sons, Inc., New York.

Fogel, M. M. and L. Duckstein. 1969. Point rainfall frequencies in convective storms. Water Resources Res. 5(6):1229-1237.

Huff, F. A. 1967. Time distribution of rainfall in heavy storms. Water Resources Res. 3(4):1007-1019.

Huff, F. A. 1968. Spatial distribution of heavy storm rainfalls in Illinois. Water Resources Res. 4(1):47-54.

McDonald, J. E. 1959. "It rained everywhere but here!" The thunderstorm encirclement illusion. Weatherwise 12:158-160, 174.

Osborn, H. B. and W. N. Reynolds. 1963. Convective storm patterns in the Southwestern United States. Bull. Int. Assoc. Sci. Hydrol. 8:73-83.

Todorovic, P. and V. Yevjevich. 1969. Stochastic process of precipitation. Colorado State University Hydrology Paper No. 35.

Wiesner, C. J. 1970. Hydrometeorology. Chapman and Hall, Ltd., London.