Journal of Hydroscience and Hydraulic Engineering
Vol. 5, No. 2, February, 1988, pp.65-83

CLUSTERING DAILY RAINFALL RECORDS PRODUCES INDEPENDENT RANDOM VARIABLES IN STOCHASTIC HYDROLOGY

By

Yasuo TAKASHIMA

Senior Engineer, EPDC International Ltd., Tokyo, Japan

SYNOPSIS

A rain storm is defined to be a sequence of rainfalls that are originated from an individual synoptic-scale meteorological disturbance, such as a typhoon, frontal wave, etc. Our basic postulation is that the occurrence of this disturbance is mutually independent and random, so is the occurrence of the rain storm. A concept of the rain cluster is introduced as the most conceivable substitute for the rain storm. Namely, the rain cluster is identified only by the uni-modal structure contained in the hyetograph of the daily rainfalls. Then, the number of occurrences, and the amount of rainfalls, of the rain cluster can be taken as the new mutually independent random variables. Consequently, well-known probability distributions in the statistics can strictly be applied for the new variables, because these distributions are based on the assumption that the variables in question should be mutually independent and random. Thus, using the rain cluster is more adequate for rigorously solving many an application problem than directly treating the daily rainfall data which are, in essence, not mutually independent nor random. Other related concepts, such as a no-rain run and a cluster interval, are also introduced. A case study concerns with the daily rainfalls observed under AMeDAS from 1975 to 1986 at Hitoyoshi, Kyushu, Japan.

INTRODUCTION

The definition of a word "independent" used in the statistics (1) is as follows. Two events A and B are independent if

$$P\{AB\} = P\{A\} \cdot P\{B\} \qquad (1)$$

where P{A} and P{B} are probabilities of the event A and event B respectively and P {AB } is the probability of the event "both A and B occur jointly." Furthermore, in this study, we classify the state of independency into two substates, an absolute independency and a relative independency. The absolute independency is illustrated by such an event that, even without resorting to the formal statistical test for independence utilizing the equation (1), one can intuitively have the correct feeling of its independency. For example, suppose a balanced coin is tossed.. For reasons of symmetry we can readily expect the events "head" and "tail" to be independent even without resorting to the formal statistical test. The equality sign of the equation (1) can, of course, be assured precisely by the absolutely independent events, if such a test is indeed performed.

An opposite concept to the absolute independency is the absolute dependency. The absolute dependency is represented concretely by such an event that, even without resorting to the formal statistical test for dependence, one can intuitively have the correct feeling of its dependency. For example, suppose that hourly rainfalls in the i-1 th day and i th day are observed. Total amount of hourly rainfalls from 0:00 to 24:00 in the i-1 th day and that in the same

hours in the next i th day are recorded as the daily rainfall of the i-1 th day and that of the ith day respectively. Suppose also that a depression, a meteorological disturbance, stagnated over the area for these two days and have originated these rainfalls. Then, since these rainfalls were originated from the same common depression, they cannot be absolutely independent but be absolutely dependent.

Next, the relative independency is an independency that can only be established by the application of certain statistical test for independence, such as the Chi-square test (2), Mann-Whitney test (3), etc., based on a certain level of significance, where the level of significance, or α, is the maximum probability of rejecting a true null hypothesis (4). The test, in essence, consists of statistical treatments of observed sample values to see whether the equality sign of the equation (1) is accepted or rejected at the given level of significance. Thus the relative independency depends on the assumed numerical value of α.

A concept of the relative dependency is an opposite concept to the relative independency. It is defined by replacing the word "independent" by "dependent" and vice versa in the above descriptions on the relative independency.

Now many statistical theories and application procedures are based on an assumption that the variable in question is an independent random variable (5). Strictly speaking, however, almost all the variables that we encounter in real world problems are absolutely dependent. Therefore, in order to cope with this reality, we must pursue either one of the following two options:

One is to search for or contrive a new random variable that is absolutely independent so that the well-developed theories and procedures in the field of the statistics can be applied strictly and conveniently. This is the line that we want to pursue in the present study, so it shall be discussed in detail later.

The other option is to admit the reality that the variable in question is, in fact, the dependent variable and to develop and construct the theory and procedure based on this reality. Many researches (6), (7) are made along this lines.

Etoh, et. al. propose a concept of a rain storm, thereby they develop very acute idea by interconnecting the statistical intermittent series (sequence of rainfall records) with an actual meteorological disturbance. According to their definition (8), the rain storm is a meso-scale meteorological disturbance such as a front, instability line, etc. The life length of the meso-scale disturbance is 10 hours or more (9).

Now from a practical view point, collecting and processing the hourly data are extremely laborious. Moreover, in some cases, the hourly data are even unavailable. (Especially in some underdeveloped countries.) Therefore we set the precondition in this study that only daily rainfall data are available. Then our definition of the rain storm should be read as follows.

A sequence of precipitations that are originated from a synoptic-scale (10) meteorological disturbance is called a rain storm, where the constituting precipitations in the rain storm are not necessarily contiguous. Short pause(s) of precipitation(s) can be intervened between the constituting precipitations in one rain storm, if these pauses occur in the period in which the sole synoptic-scale disturbance prevails.

According to the definition (10) given in the meteorology, the synoptic-scale disturbance includes such meteorological phenomena as migratory anticyclone, migratory cyclone, typhoon, frontal wave, etc. The wave length of the disturbance in horizontal direction ranges from 1000 km to 5000 km, the vertical height is about 10 km, and the life span varies between one day and one week. It is noted that above definition is a conceptual one and that the stated numerical values should not be taken as strict quantitative criteria but be interpreted as rough illustrations of the standard size of the synoptic-scale disturbance. This implies that some ambiguities and difficulties are encoutered when the rain storm is identified from observed meteorological data.

Now it is obvious that an occurrence of a rain storm is absolutely independent and random because each synoptic-scale disturbance which originates the rain storm occurs independently and randomly. (This is our basic postulation in this study.) It is important to note, therefore, that because of this absolute independency, the occurrence of rain storm can be regarded as independent and random even without applying any formal test for its independency.

Next problem is how to extract and compile the rain storm data. Meteorological agencies publish regularly daily synoptic charts (weather maps), monthly weather reports, etc. These contain the information on the synoptic-scale disturbance, hence they are especially suited for our purpose. (Incidentally, this is one of the reasons that we adopt the synoptic-scale disturbance as the base of the definition of the rain storm.)

However, the work involved in the extraction and compilation is prohibitively laborious because the work is not computer oriented but calls for human's judging ability to solve the ambiguities and difficulties exist in the information. Therefore, we must contrive somewhat expedient means to substitute for the rain storm by using the daily rainfall records directly without resorting to the various information above mentioned. Two substitutes, a rain run and rain cluster, will be discussed in the following sections.

RAIN RUN AND NO-RAIN RUN

The first substitute that is considered is called a rain run, where a run is the borrowed term from the field of the statistics and its definition reads(11): "In any ordered sequence of elements of two kinds, each maximal subsequence of elements of like kind is called a run." For example, the sequence

$$\underline{r\,r}\;\underline{o\,o}\;\underline{r\,r\,r\,r}\;\underline{o}\;\underline{r}\;\underline{o}\;\underline{r\,r}\;\underline{o} \qquad (2)$$

opens with an r run of length 2; it is followed by runs of length 2,4,1,1,1,3,1, respectively. A length of a run is the number of elements contained in the run. In our case, the "element of one kind" corresponds to the rain day denoted by r in the expression (2), and the "element of another kind" is the no-rain day denoted by o. Thus the rain run can be a substitute of the rain storm because each rain run is separated by each no-rain run and hence it is likely that each run corresponds with each synoptic-scale disturbance.

Now in order to proceed the discussion further, it seems necessary to illustrate an actual example, which will be presented in the following. We will take up the daily rainfalls at Hitoyoshi, Kumamoto prefecture, Kyushu, Japan, as our example. Among all the uninterruptedly observed data at Hitoyoshi Weather Station (15) for more than 40 years, only a fraction of them, that is, from the 1st January, 1975 to 31st December, 1986 is extracted for this study.(Hitoyoshi Weather Station belongs to Kumamoto Meteorological Observatory, Japan Meteorological Agency.) This period is selected by considering the fact that in around 1975 a new meteorological observation network over Japan, named "Automated Meteorological Data Acquisition System (AMeDAS)" was set up by Japan Meteorological Agency and the observation network has greatly been strengthened since then. (Of course, Hitoyoshi Weather Station forms a link in the network.) Since, even for this period (1975 - 1986, 12 years), the whole daily data are so voluminous, their complete presentation here is omitted (16), and only a part of them, from the 1st to the 30th April, 1975 is shown on Fig. 1 for illustration. (The complete data are included in the monthly reports (12), issued by Kumamoto Meteorological Observatory.)

Fig. 1(a) illustrates the grouping of the daily rainfalls into rain runs as introduced at the beginning of this section, and Fig.1(b) lists the amounts of daily rainfalls observed at Hitoyoshi. (Fig. 1 (c) will be explained in the

later section.) By means of Fig.1 we can easily see the one-to-one correspondences of the rain runs with their respective meteorological phenomena. Several interesting points will be described below.

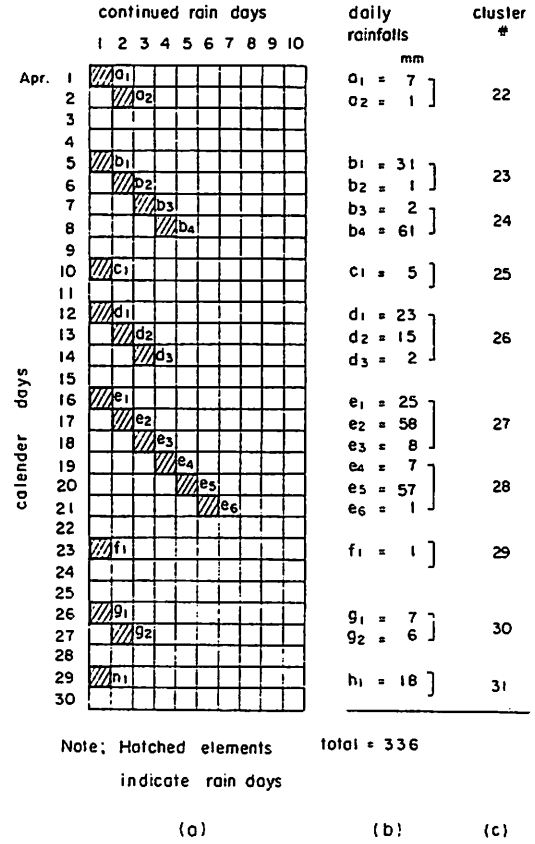

Fig. 1 Daily rainfalls at Hitoyoshi, 1st - 30th, April, 1975

For a few days at the beginning of April, the pressure pattern over the southern part of the Kyushu district being that of a winter type (cold high pressure moved onto the area from the Asian Continent built up the so-called west-high-east-low pressure pattern), it snowed especially in mountaneous zones. Hence the cause of the rain run $\{a_1\ a_2\}$ was the winter type pressure pattern. On the 5th, it rained, due to a passing of a pressure trough over the area. On the 6th, it was fine (actually b_2 = 1mm), covered by a moving high pressure. On the 7th and 8th, it rained due to a passing of a pressure trough. Thus we see that, although we have classified the sequence b_1, b_2, b_3, b_4 into the single rain run, rigorously it shuld be divided into two rain storms, either $\{b_1\ b_2\}$ and $\{b_3\ b_4\}$ or $\{b_1\}$ and $\{b_2\ b_3\ b_4\}$. This ambiguity illustrates the nature of the rain run as being the second (or the third) best version for representing the rain storm. According to the reports, 14th was fine covered by a moving high pressure. But actually there was a rain (d_3 = 2mm) recorded at Hitoyoshi. The monthly reports generally describe the synoptic-scale meteorological disturbances all over the southern part of the Kyushu district, while the rainfall record d_3 is a point rainfall observed at Hitoyoshi. Therefore, we interprete the above discrepancy as follows: On the whole, it was fine on the 14th, but there was a local rainfall at Hitoyoshi due to small-scale disturbance or convection. This is one of the examples of the difficulties to classify all the daily point rainfalls exhaustively by referring to the descriptions of the monthly reports. According to our definition of the rain run, the rainfall d_3 is an element of the rain run $\{d_1\ d_2\ d_3\}$ and this may be reasonable if we interprete the d_3 as a consequence or aftermath of the low pressure, which is a synoptic-scale disturbance, passed on the 12th and 13th.

From the 15th through 20th, a stationary front stagnated over the sea south to Kyushu district. Moreover, several low pressures crossed (or superposed) over the front from west to east one after another, orginated rains, e_1,e_2, e_3, e_4, e_5, and e_6. Especially on the 16th and 17th, the front moved up toward the middle part of Kyushu district and was stimulated by low pressure passing over it, and so, heavy rainfall (e_2 = 58mm) was recorded. Here we have another example of difficulties. The cause of this heavy rainfall was the stationary

front plus the low pressure, that is, a compound cause. But other rainfalls might have been originated either by the compound cause or by the single cause (i.e. the stationary front only). Thus subdividing the rain run { e_1 e_2 e_3 e_4 e_5 e_6 } into several rain storms is difficult, and hence classifying these rainfalls into one rain run is somewhat unreasonable in this case.

From 24th to 26th, it was fine covered by a moving high pressure, but from the evening of the 26th to the morning of the 27th, it rained (g_1 = 7mm, g_2 = 6mm), due to a passing of a pressure trough. This is an example that the daily rainfalls g_1 and g_2 do not constitute a mutually independent random variable, but the rain run { g_1 g_2 } does.

So far, we have focused our attention on the extraction of the rain runs. However, we can also extract the no-rain runs as well. There are altogether 7 no-rain runs, that is, { 3rd, 4th }, { 9th }, { 11th }, { 15th }, { 22nd }, { 24th, 25th } and { 28th } in April, 1975 as can obviously be identified on Fig. 1 (See the footnote of Table 2 for the reason why the 30th is not included in the no-rain runs in April).

Thus, we have seen the mutually independent and random nature of the rain runs and no-rain runs and so we confirm that the rain run can, in fact, be a substitute for the rain storm.

Now, therefore, we form all the rain runs and no-rain runs for whole the period from 1975 to 1986. The result are summarized as follows. Within this period, total number of calendar days is 4,383 days of which the number of rain days and the number of no-rain days are 1,616 days and 2,767 days respectively. The total number of the rain runs and the total number of no-rain runs are both 782 as shown on Table 1. On the average, the length of the rain run is 1,616/782 = 2.1 days and that of the no-rain run is 2,767/782 = 3.5 days. The mean rainfall per rain run is 35.5mm.

Although we obtained much more information on the rain run, its complete presentation here is omitted, because the rain run is not preferable to another substitute, a rain cluster, as will be discussed in the following section.

Table 1 Number of rain runs and no-rain runs
Jan. 1, 1975 - Dec. 31, 1986, Hitoyoshi

Length in days	Number of rain runs		Number of no-rain runs
1	349		230
2	239		166
3	100		120
4	44	773	65
5	29		48
6	7		41
7	5		26
8	3		25
9	1		12
10	2	9	12
11	0		13
12	2		4
13	1		8
14	-		4
15	-		3
16	-		3
17	-		1
28	-		1
Total	782		782

RAIN CLUSTER

In the preceding section, we defined the rain run and adopted it as a substitute for the rain storm. We saw, in the actual example, that the maximal length of the rain run was 13 days and there were 9 rain runs (out of total 782 rain runs) whose lengths were larger than 7 days (See Table 1). The lengths of these 9 rain runs are somewhat larger than the standard length (from 1 day to 1 week) of the rain storm (synoptic-scale disturbance) given in the introductory section. The reason why these larger lengths come about is found in the definition and the way of forming the rain run. Namely, as long as rain days continue, they are collected together and are formed into a single rain run without any regard to the underlying meteorological phenomena that these rainfalls might actually be originated from two or more different synoptic-scale disturbances. In fact, in the preceding example we saw that the single rain run $\{ b_1\ b_2\ b_3\ b_4\}$ should be separated into two rain storms. From the results, therefore, we see that the length of the rain run always gives the upper bound of the conceivable range of the length of the rain storm.

Now therefore the problem is to select more adequate substitute for the rain storm. However, the difficulty for the selection of the perfect substitute remains unchanged. Therefore, our strategy this time is that we will select such a substitute that gives the lower bound of the length. Then, the real length will, anyway, be in-between them.

Etoh et. al. (7) proposed an energy relaxation process of a sequence of concentrated energy bursts (i.e. sequence of points) as a model for the rainfall time series. Since these bursts relax during ensuing time period, there are as many relaxing time periods as there are energy bursts. Thus if some of the adjacent relaxing periods overlap each other, they are merged into one state of the process. Then, the problem here is how to count the number of energy bursts from the observed rainfall records, because the actual records represent only the number of merged states, not the number of energy bursts. The practical method, however, is not clear to count this number on which the probability of the occurrences of energy bursts stands.

Now, our second substitute, a rain cluster, is contrived and formed as follows. We know from our experiences that whenever a rainfall occurs, the intensity of the rainfall is relatively low at the beginning, it then goes up to its maximum and decreases toward the end of the rainfall. We also know that this rainfall is originated from a certain meteorological cause, e.g. a passing of a low pressure, etc. Note that we are not concerned with hourly (nor minutely) rainfalls but are interested in the daily rainfalls. Although the intensities of the former oscilate so that often they form many peaks and bottoms on an hourly hyetograph, the daily rainfall as the sum of the hourly rainfalls does not fluctuate so often as does the hourly rainfall. Rather, within a period of a few days over which a sole meteorological disturbance prevails, the shape of hyetograph of daily rainfalls is usually simple, i.e. mono-peaked type or uni-modal. Therefore it is reasonable to infer that, in most cases, a meteorological disturbance corresponds to such a group of daily rainfalls that the shape of the hyetegraph is uni-modal.

Now we define the rain cluster as follows. If a uni-modal structure is formed by a portion of the rain run, that portion forms a rain cluster, and if a uni-modal structure is formed by the whole single rain run, that rain run forms a rain cluster.

For example, Fig. 2(a) shows a hyetograph from 11th to 15th April, 1975 at Hitoyoshi (based on the data given in Fig. 1). The uni-modality of the hyetograph is clearly seen on this graph. Hence, the group of the daily rainfalls from the 12th to 14th is identified as the rain cluster according to the definition. Incidentally, this group of rainfalls (see $\{ d_1, d_2, d_3\}$ in Fig. 1) was originated from a passing of a low pressure over the sea south to Kyushu district. Thus, even without resorting to the monthly meteorological reports,

we have obtained, in this case, the exact result that this rain cluster coincides with the distinct rain storm.

The second example concerns with the daily rainfalls from the 15th to 22nd April, 1975. (These rainfalls were already classified as the rain run $\{e_1, e_2, e_3, e_4, e_5, e_6\}$ in Fig. 1.) The hyetograph of these rainfalls is plotted on Fig. 2(b). It starts with the positive slope followed by the negative slope forming the first peak at the 17th. The slope then turns up from negative to positive at the 19th and the second peak is formed at the 20th. After that, the slope is negative again to the end of the hyetograph. Hence these daily rainfalls should be separated into two rain clusters because the shape of the hyetograph clearly represents the bimodal structure. The boundary day between these separated rain clusters is the 19th where the sign of the slope changes from negative to positive.

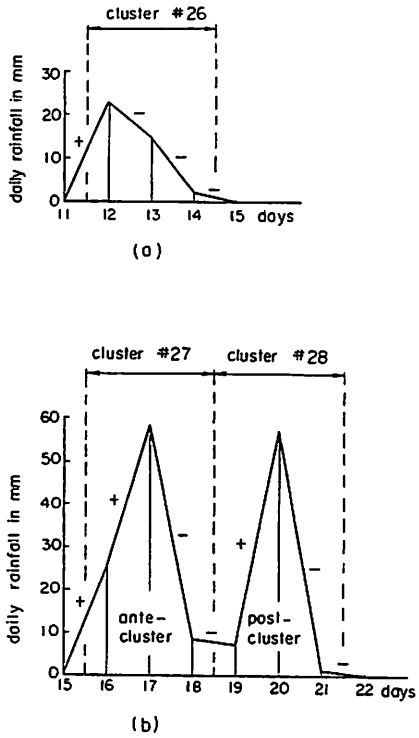

Fig. 2 Identification of rain clusters, Apr. 1975, Hitoyoshi

Generalizing these observations, we establish a separation criterion as follows: Whenever a change of the sign of slope of hyetograph from negative to positive occurs, those daily rainfalls that constitute the hyetograph should be separated into two distinct rain clusters. This criterion, however, is not a complete one. For example, it fails to specify that to which separated cluster the boundary day itself should belong. For the sake of completeness and uniquenness, therefore, we must add several supplementary rules to the criterion.

However, since all of those supplementary rules are not theoretical ones but are merely for convention to obtain the unique result, the detailed explanation is omitted here except that the results of their application are exhibited on Fig. 1 (c).

Anyhow we have seen by way of the examples that the criteria above specified have worked reasonably well and that even if we do not resort to the meteorological reports, the identification of the rain clusters is seemingly possible.

Finally, in connection with the rain cluster, we will define the cluster interval as follows. The time interval (or distance on the time axis) between the centers of gravity of the adjacent rain clusters is called the cluster interval. One of the interesting properties of the cluster interval is that it has a theoretical (or proper) distribution, i.e. an exponential distribution. We will take up this matter shortly.

So far we have defined and dicussed the identification of the rain clusters, no-rain runs and cluster intervals by observing the basic requirement that the variables should be absolutely indenpendent. Now, therefore, we are in

a situation that we can develop proper probability distributions that represent these variables formally. For example, since the occurrences of the rain cluster can be compared favorably with the well-known Bernoulli trials (13), many theories established in the field of statistics can be applied rigorously to our variables. Thus, the exponential distribution (14) can properly be assigned for the cluster interval, because the cluster interval (i.e. time interval of the cluster occurrences) is comparable with the waiting time for the success in the sequence of the Bernoulli trials and this waiting time is proved to be exponentially distributed.

Unfortunately, however, we are not given similarly well established distributions for other variables. Therefore, we select the ones for the length of rain cluster, the length of no-rain run and the amount of cluster rainfall by a trial-and-error method. Namely, we select them from the inventory of the distributions available in any handbook (15) of the statistics (we call these distributions the standard distributions) employing a numerical fitting procedure, where the inventory includes such distributions as Poisson, geometric, exponential, gamma, etc.

Note that all the standard distributions in the inventory are based on the assumption that the variable in question should be the mutually independent random variable. Since our variables are all mutually independent, we are equipped with power to select any of these distributions in the inventory. More concrete details on the selection, including such manipulations as compounding, truncation of the distributions will be illustrated in the following case study.

A CASE STUDY

This case study uses the same data as were introduced in the previous section, that is, the daily rainfall data from Jan. 1, 1975 to Dec., 31, 1986 at Hitoyoshi. All the rain clusters in this period were identified by

Table 2 Rain clusters and no-rain runs from 1st to 29th, April 1975

No-rain run				Rain cluster									
No.	Start day		Length in days	No.	Start day		Length in days	Amount of rainfall in mm					Location of center of gravity in day No.
	Start day #	Date			Start day #	Date		1st day	2nd day	3rd day	4th day	Total	
				22	91	Apr. 1	2	7	1	-	-	8	91.1
20	93	Apr. 3	2										
				23	95	Apr. 5	2	31	1	-	-	32	95.0
				24	97	Apr. 7	2	2	61	-	-	63	98.0
21	99	Apr. 9	1										
				25	100	Apr. 10	1	5	-	-	-	5	100.0
22	101	Apr. 11	1										
				26	102	Apr. 12	3	23	15	2	-	40	102.5
23	105	Apr. 15	1										
				27	106	Apr. 16	3	25	58	8	-	91	106.8
				28	109	Apr. 19	3	7	57	1	-	65	109.9
24	112	Apr. 22	1										
				29	113	Apr. 23	1	1	-	-	-	1	113.0
25	114	Apr. 24	2										
				30	116	Apr. 26	2	7	6	-	-	13	116.5
26	118	Apr. 28	1										
				31	119	Apr. 29	1	18	-	-	-	18	119.0
(7)			(9)	(10)			(20)					(336)	

Note: 1) Day # indicates serial day number starting on 1st Jan. 1975 (day #1) & ending on 31st Dec. 1986 (day #4383).

2) From the 30th Apr. to the 3rd May, no-rains. Hence the 30th Apr. is included in the no-rain run #27 which is a no-rain run in May, and so it is not included in this Table.

applying the criteria discussed in the preceding section. Since the whole results are so voluminous, only a part of them is excerpted and illustrated on Table 2, which contains the rain clusters and no-rain runs from the 1st to 29th April 1975 only.

Some of the main results for the whole 12 years are as follows. The total number of rain clusters is 887 and the length of rain cluster ranges from 1 day to 7 days. The mean length and standard deviation are calculated to be 1.8 days and 0.9 day respectively. (see Table 3.) The last column in Table 3, the fitted probability distribution, will be explained later in this section.

Table 3 Lengths of rain clusters, Jan. 1975 - Dec. 1986, Hitoyoshi

Length of a rain cluster in days	Number of rain clusters	Observed frequency	Fitted probability distribution
1	394	.444	.473
2	322	.363	.319
3	123	.139	.143
4	35	.040	.048
5	11	.012	.013
6	1	.001	.003
7	1	.001	.001
Total:	887	1.000	1.000

Mean length: 1.821 days
Standard deviation = 0.931 day
Fitted distribution = Truncated Poisson distribution (See Fig. 3)

Table 4 Number of rain clusters by month and by year, Jan. 1, 1975 - Dec. 31, 1986, Hitoyoshi

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	TOTAL
1975	9	7	5	10	8	6	7	7	7	6	4	5	81
1976	6	7	6	10	8	6	7	7	5	6	6	6	80
1977	4	5	9	7	7	7	5	7	8	1	6	7	73
1978	8	5	6	8	4	6	8	7	5	4	3	5	69
1979	6	7	8	6	5	7	7	7	7	1	6	7	74
1980	7	3	8	9	6	7	8	8	3	4	4	7	74
1981	4	7	9	8	6	5	9	6	5	7	5	3	74
1982	6	6	8	8	6	5	5	6	6	4	6	5	71
1983	4	4	8	9	6	7	7	5	7	6	6	3	72
1984	6	6	8	6	6	9	7	4	7	5	2	5	71
1985	7	7	9	7	6	7	5	4	5	5	6	7	75
1986	5	4	7	6	5	6	9	8	6	6	6	5	73
TOTAL	72	68	91	94	73	78	84	76	71	55	60	65	887
MEAN	6.0	5.7	7.6	7.8	6.1	6.5	7.0	6.3	5.9	4.6	5.0	5.4	6.2
STD. DEV.	1.6	1.4	1.3	1.5	1.2	1.1	1.4	1.4	1.4	1.9	1.4	1.4	1.6
COEFF. OF VAR	0.27	0.25	0.17	0.19	0.19	0.17	0.20	0.22	0.23	0.42	0.28	0.27	0.27

It should be noticed that the range of the length of rain cluster (i.e. from 1 to 7 days) just coincides with that given in the introductory section for the length of the synoptic-scale rain storm. This implies that our scheme of contrivance, the rain cluster, as a substitute for the rain storm, has worked well. We prefer, therefore, the rain cluster to the rain run and adopt the rain cluster as the substitute for the rain storm in the following discussion.

The break-down of the number of occurrences of the rain clusters into month and year is shown on Table 4 together with some statistics computed. It is seen from this table that the mean number of occurrences of the rain clusters is minimal (i.e. 4.6) in October and is maximal (i.e. 7.8) in April. Annual average is 6.2 clusters per month. On the other hand, the amount of rainfall per one cluster ranges widely from 1mm to 727mm. (see Table 5) Incidentally, the reason why the amount of rainfall less than 1mm is excluded is as follows. Robotto rain gauges of the AMeDAS transmit the signals of rainfalls at every 1mm. As the result, the minimum amount of rainfall recorded is 1mm and the rainfalls whose amounts are less than 1mm are not observed.

Table 5 Amount of rainfall of rain clusters 1975 - 1986, Hitoyoshi

Cell #	Upper boundary of each cell in mm	Number of clusters	Observed frequency
1	67	784	.88
2	133	71	.08
3	199	17	.02
4	265	10	.01
5	331	3	.00
6	397	1	.00
7	463	0	.00
8	529	0	.00
9	595	0	.00
10	661	0	.00
11	727	1	.00
Total:		887	1.00

Mean rainfall per cluster = 31.268 mm
Standard deviation = 48.357 mm

Now, the variation or distribution of observed lengths and amounts of rainfalls of the rain clusters are best visualized by creating histograms. These are shown on Fig. 3 and Fig. 4. On these histograms, thick bars (thick spikes) in Fig. 3 and cells in Fig. 4 represent the observed distributions. (Each block of equal width which consitutes the histogram is called a cell.) The thin spikes in Fig. 3 and a chained curve in Fig. 4 show the fitted probability distribution which will be discussed soon.

The no-rain runs for the same period (1975 - 1986) are also extracted. The results are shown on Table 6 and Table 7. It is seen that the length of no-rain run varies from the minimum 1 day to the maximum 28 days. The mean length is 3.5 days and the mean number of occurrences of the no-rain run is 5.4 runs per month. The histogram of the length of the no-rain runs is also created as shown on Fig. 5.

Finally, interval lengths (i.e. distances on the time axis) between centeres of gravity of the rain clusters are computed. The result is that the total number of intervals is 886, the maximum interval length is 29.0 days, the minimum interval length is 1.5 days and the mean interval length is 4.9 days. The histogram of the interval length is also created (see Fig. 6). The fitted curve appearing in Fig. 6 will be explained soon.

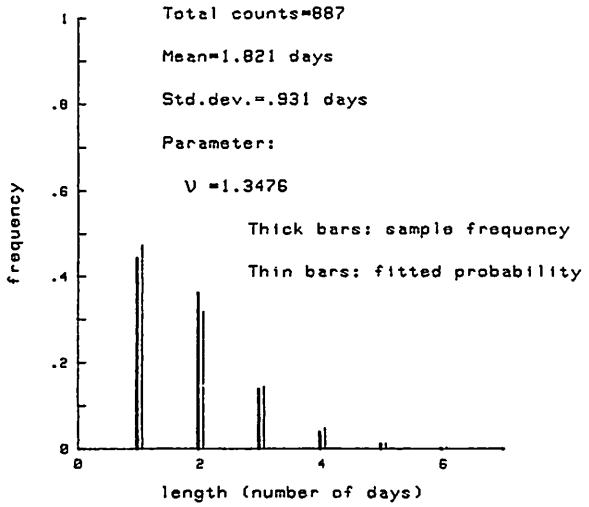

Fig. 3 Number of rain days fitted truncated Poisson distribution, Hitoyoshi.

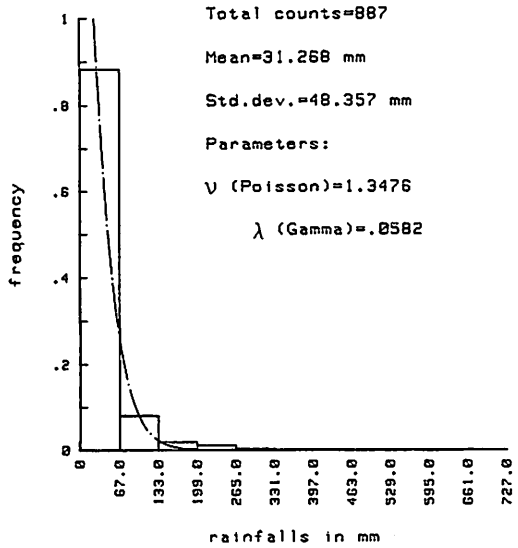

Fig. 4 Histogram of rainfalls & fitted gamma & truncated Poisson compound, Hitoyoshi.

A few comments may be in order at this point. Recall that all the data used in this case study are daily observation records and hence the minimum cluster length, the minimum length of no-rain run and the minimum length of cluster interval are all 1 day. This implies that the range of the variables (i.e. the lengths in days) less than 1 day is completely eliminated from the sample space. However, the range of the variable of the standard distribution function is either from 0 to ∞ (or 0, 1, 2, ...) or from $-\infty$ to ∞ (or ..., -1, 0, 1, ...).

Table 6 Lengths of no-rain runs, Jan. 1975 - Dec. 1986, Hitoyoshi

Length of a rain cluster in days	Number of rain clusters		Observed frequency	Fitted probability distribution
1	230			
2	166	516	.660	.622
3	120			
4	65	113	.144	.180
5	48			
6	41			
7	26	92	.118	.123
8	25			
9	12	24	.031	.036
10	12			
11	13			
12	4	25	.032	.024
13	8			
14	4	7	.009	.007
15	3			
16	3			
17	1	4	.005	.005
18	0			
19	0	0	0	.001
20	0			
21	0			
22	0	0	0	.001
23	0			
24	0	0	0	.000
25	0			
26	0			
27	0	1	.001	.000
28	1			
	782		1.000	1.000

Mean length: 3.588 days
Standard deviation = 3.149 days
Fitted distribution = Truncated Poisson distribution (See Fig. 8)

For example, the range of the variable x of the standard exponential distribution function is from 0 to ∞ as can be seen from the following formula,

$$f(x) = \alpha e^{-\alpha x}, \quad x \geq 0 \tag{3}$$

where α is a parameter to be estimated from the observed data. However, if the range of x is bounded away from x_0 ($x_0 > 0$), the density (3) shall be adjusted to conform with the new range. Thus, the truncated exponential density,

$$f_t(x) = \alpha e^{-\alpha|x - x_0|}, \quad x \geq x_0 > 0 \tag{4}$$

where,
x_0 : truncation point (x_0 = 1 day in this case)
should be adopted for the truncated variable. In fact this truncated exponen-

Table 7 Number of no-rain runs by month and by year, Jan. 1, 1975 - Dec. 31, 1986, Hitoyoshi

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	TOTAL
1975	9	5	5	7	7	3	6	8	3	5	5	3	66
1976	6	5	6	8	7	4	4	7	5	5	7	4	68
1977	4	6	8	7	7	5	6	7	5	2	6	4	67
1978	8	4	6	6	5	5	7	6	5	4	4	4	64
1979	7	7	7	5	5	4	5	4	6	2	5	7	64
1980	7	3	7	8	7	3	4	6	3	4	5	7	64
1981	4	6	6	7	5	4	8	6	5	6	4	3	64
1982	5	6	8	8	4	6	3	6	5	4	5	5	65
1983	4	4	7	8	5	7	5	4	4	7	5	4	64
1984	5	6	8	6	4	6	5	5	5	5	3	5	63
1985	6	7	9	6	7	6	4	2	3	5	7	6	68
1986	5	4	7	7	4	4	5	6	7	5	6	5	65
TOTAL	70	63	84	83	67	57	62	67	56	54	62	57	782
MEAN	5.8	5.3	7.0	6.9	5.6	4.8	5.2	5.6	4.7	4.5	5.2	4.8	5.4
STD. DEV.	1.6	1.3	1.1	1.0	1.3	1.3	1.4	1.6	1.2	1.4	1.2	1.4	1.5
COEFF. OF VAR	0.28	0.25	0.16	0.14	0.24	0.27	0.27	0.29	0.26	0.32	0.23	0.29	0.28

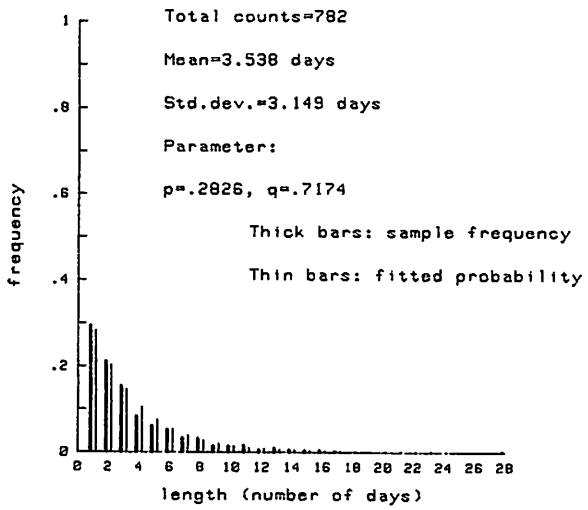

Fig. 5 Number of no-rain days in no-rain runs & fitted truncated geometric distribution, Hitoyoshi

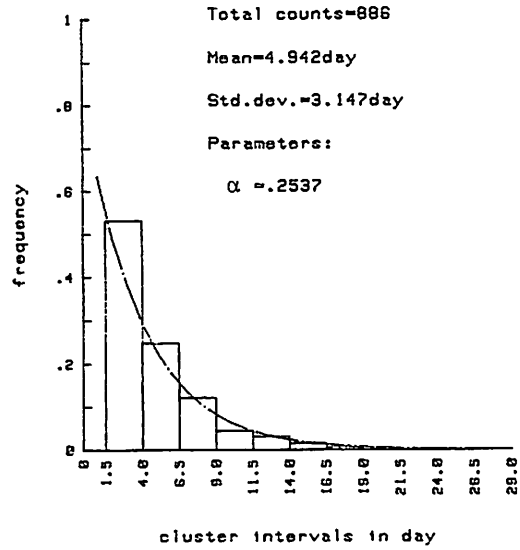

Fig. 6 Histogram of cluster intervals (c. to c.) & fitted truncated exponential density, Hitoyoshi

tial density is applied for the interval length of the rain clusters as will be seen later.

A similar comment is applicable for the distribution of the amount of the rain cluster, because the observation of the amount of the rainfall below 1mm is truncated under the AMeDAS as explained already. However, in this case, the amount of 1mm rainfall is small compared with the mean amount (31.3mm, see Table 5) of the cluster rainfall, the truncation may not be so indispensable.

Now the selected distributions are exhibited in the following.

(i) For the length of rain clusters, the truncated Poisson distribution is selected, and its distribution and mean are expressed in the forms,

$$p(n) = \frac{1}{1-e^{-\nu}} \frac{e^{-\nu}\nu^n}{n!}, \quad n = 1, 2, \cdots \tag{5}$$

$$m_n = \sum_{n=1}^{\infty} n\, p(n) = \frac{\nu}{1-e^{-\nu}} \tag{6}$$

Since the value m_n is given by the observed data (m_n = 1.821 days), the parameter ν can be estimated by this equation (ν = 1.3476 days). Consequently, the distribution can be calculated by (5) for n = 1, 2, ... The thin spikes on Fig. 3 shows thus calculated distribution.

(ii) For the amount of rainfalls of rain clusters, the truncated Poisson and gamma compound density is selected, which is expressed in the forms,

$$h(t) = \sum_{n=1}^{\infty} \frac{\lambda e^{-\lambda t}}{(n-1)!}(\lambda t)^{n-1} \frac{1}{1-e^{-\nu}} \frac{e^{-\nu}\nu^n}{n!}, \quad t>0 \tag{7}$$

$$m_t = \frac{m_n}{\lambda} \tag{8}$$

where, the value m_n represents the same value as defined in (6) and so, its value has been obtained already. Since the value m_t is calculated from the observed data (m_t = 31.268 mm), the value of parameter λ is estimated using the equation (8). (λ = 0.0582). Consequently, the density can be calculated by (7) for $t > 0$. (the value of ν has already been obtained by (6) to be 1.3476). Since the density (7) is a continuous in t, it is plotted as a curve (a chain line) on Fig. 4.

(iii) For the length of no-rain runs, the truncated geometric distribution is selected, and its distribution expressed in the forms,

$$P(k) = p(1-p)^{k-1}, \quad k = 1, 2, \cdots \tag{9}$$

$$m_k = \frac{1}{p} \tag{10}$$

Since the value m_k is given by the observed data (m_k = 3.538 days), the parameter p can be estimated by this equation. (p= 0.2826) Consequently, the distribution can be calculated by (9) for each of k = 1, 2, ... The results are plotted by thin spikes on Fig. 5, together with the observed distribution (thick spikes) of no-rain runs.

(iv) For the interval length of the rain clusters, the truncated exponential density is selected. This density has been discussed already as the density (4). The mean m_x of x is obtained by

$$m_x = \frac{x_0 + \alpha}{\alpha} \tag{11}$$

Since the value m_x is given by the observed data (m_x = 4.942 day), the value of parameter α can be estimated from this equation (α = 0.2537). Consequently, the density can be calculated by (4). Since it is continuous in t, it is plotted as a curve (a chain line) on Fig. 6.

SUPPLEMENTAL COMMENTS

The first supplemental comment is as follows. In the preceding section we set up the rule and criteria to identify the rain cluster from the daily rainfall records observed at one place. However, at this stage of study, we try to see the rain cluster with a bit broadened eyes. Namely, although our discussion has so far been based only on the point rainfall data, now we take areal rainfall data into our consideration.

Suppose that several rain gauges are scattered over a basin and the daily rainfalls are observed. By using all of these scattered observations, the correlation coefficient between the daily rainfalls of a certain day and the subsequent day can easily be calculated 16). The value of the correlation coefficient thus calculated indicates the closeness of the relation between these two daily rainfalls. In other words, it gives us an idea how similar these two rainfalls are. This suggests us that the correlation coefficient can be used as a base of the extraction rule of the rain cluster instead of the sign change of the slope of hyetograph as previously adopted. It is obvious that multiple point observation data are more effectual than single point observation data for

obtaining the information on the synoptic-scale disturbance.

Moreover, by utilizing these multi-point observation data and the newly obtained rain clusters, we may estimate more reliable mean areal rainfall over the basin which, up to present, has been estimated only by taking simple arithmetric mean of the available data regardless how each of the observed data shares the role to represent the true basin mean. However, this kind of study is detoured in this paper and suggested here as one of the application fields of the rain cluster.

The second supplemental comment is as follows. In the foregoing study we formed the rain cluster from the daily rainfall data. But it may sometime be needed to predict the daily rainfall inversely from the rain cluster.

In this connection, Nagao 17) proposed a method to estimate the rainfall of shorter time interval (hourly rainfall, say) from the given data of rainfall of longer time interval (daily rainfall, say). Although the rain cluster is a bit different from the rainfall of this longer time interval, Nagao's method and idea will be of reference for us.

Since the daily rainfalls within a rain cluster are not mutually independent, we cannot directly estimate the probability distribution of the daily rainfall by selecting and/or manipulating some of the standard distributions. One of the ways to overcome this difficulty is to introduce the auto-correlation coefficient of the daily rainfalls based on the reality that the daily rainfalls are, in fact, absolutely dependent. This is the way that Nagao adopted.

But we will pursue another way in this study, i.e. we want to find such measures of rainfalls that are derived from mutually independent random variables. Recall Fig. 1 and Table 2 and suppose we extract a sample of rainfalls $\{ a_1, b_1, b_3, c_1, d_1, \ldots \}$. Then the constituent daily rainfalls are obviously independent because the rain clusters $\{ a_1\ a_2 \}$, $\{ b_1\ b_2 \}$, $\{ b_3\ b_4 \}$, $\{ c_1 \}$, $\{ d_1\ d_2\ d_3 \}$, ... are mutually independent. Similarly, another sample of rainfalls $\{ a_2, b_2, b_4, d_2, \ldots \}$, say, can be extracted which constitutes another sample of the mutually independent random variable.

We generalize the above procedures as follows. Suppose that all the n th day (n=1,2,..., N) rainfalls in the rain clusters are extracted. Then the extracted daily rainfalls constitute a sample of the independent random variable. Thus we can obtain altogether N samples of the daily rainfall data that are independent random variables.

As has been stressed, any standard distribution should not be applied for the aggregate of daily rainfall data, but now it can be applied for each of these samples of daily rainfalls. On Fig. 7 are shown the histograms of the n th day rainfalls, n=1,2,3,4 of the rain clusters formed in the foregoing case study. The shapes of the histograms remind us of highly skewed distributions, such as gamma, Poisson and gamma compound, etc. It is noted that, since the number of rain clusters (i.e. number of counts) whose lengths are more than 4 days is small (see Table 3), the histograms for the 5th day, 6th day and the 7th day rainfalls are not created in this case.

Based on these histograms, it is possible to select the most appropriate distribution for each of the samples of the daily rainfalls above extracted. Then, by gathering all the results, we will have a medley of the daily rainfall distributions, the distribution of the 1st day rainfall, the distribution of the 2nd day rainfall, ..., in the rain cluster. Thus if we shall predict a daily rainfall, the prediction shall be done in two steps. In the first step we shall predict whole rain clusters using the distribution of the rain cluster already estimated. Then in the next step, we shall predict the daily rainfall of the nth day in the rain cluster utilizing the nth day rainfall distribution above suggested, for n=1,2, ..., N. By employing these two step procedure, we shall predict more reliable daily rainfalls because we can apply more authentic probability distributions. However, we will not proceed this matter any further in this study. Here we are only to illustrate the usefulness of the rain cluster and to suggest some of the promisable application fields.

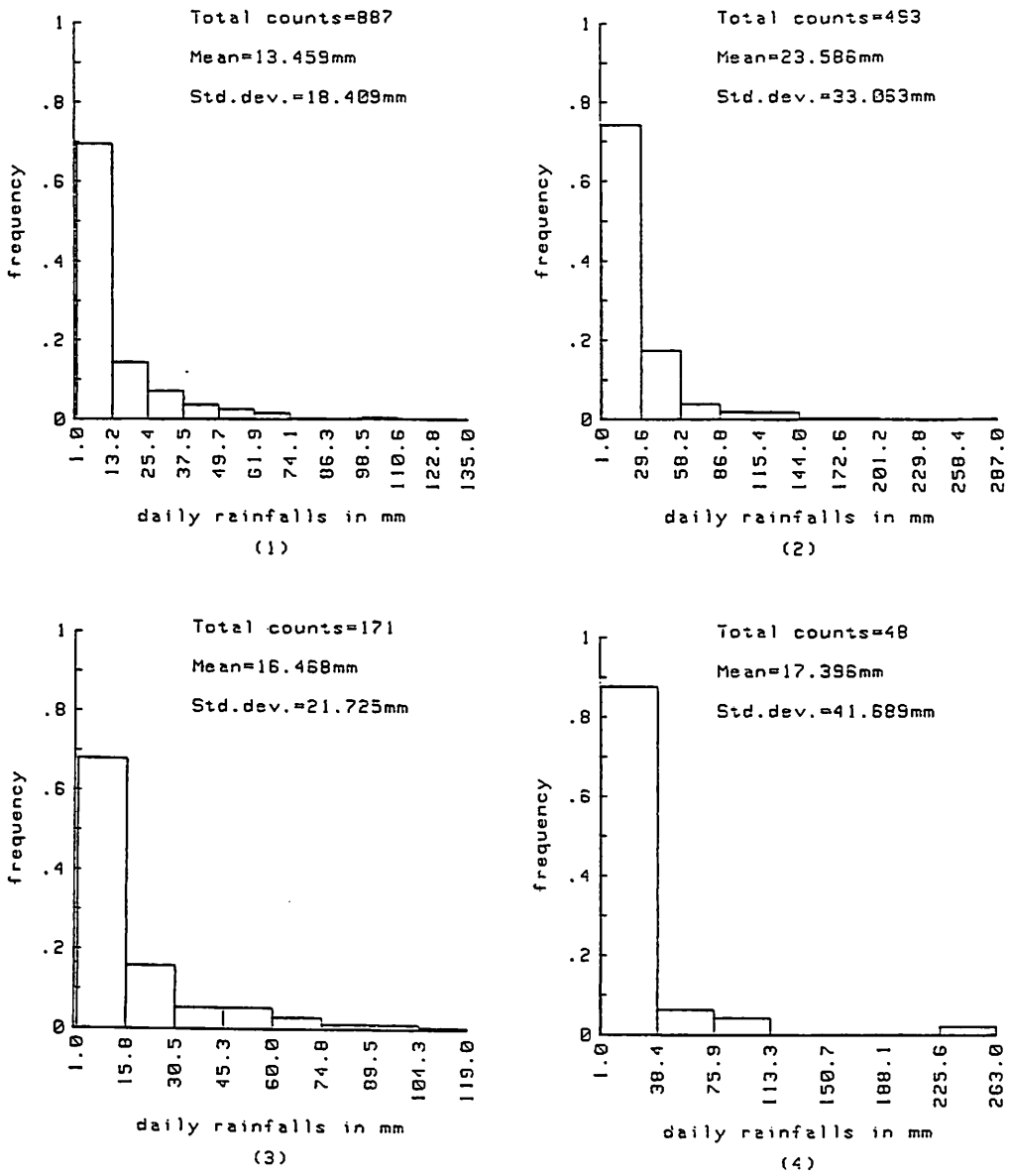

Fig. 7 Histograms of n th day rainfalls in rain clusters, n = 1,2,3,4, Hitoyoshi.

SUMMARY AND CONCLUSION

The main purpose of this study is to establish new mutually independent random variables that can be derived from daily rainfall records. Basically, the postulate of this study is that the occurrences of the synoptic-scale meteorological disturbances, such as typhoon, frontal wave, etc. whose life lengths vary from one day to one week, are random and mutually independent.

Based on this postulate, the key idea is to let the new variable correspond to the synoptic-scale disturbance one by one. By doing so, the new variable established is obviously the mutually independent random variable.

A group of daily rainfalls originated by an individual synoptic-scale meteorological disturbance is called a rain storm. In order to facilitate the actual works involved in identifying all the rain storms from the daily rainfall records, two substitues for the rain storm are contrived, the rain run and the rain cluster. They were compared by means of the case study and it is concluded that the rain cluster is preferable to the rain run. A notion of no-rain run is also introduced. Then, by combining all the rain clusters and no-rain runs, whole daily rainfall data are partitioned into a sequence of alternations of the rain clusters and no-rain runs. Finally, the cluster interval is defined as the time interval between the centers of gravity of the adjacent rain clusters. Since the occurrence of the rain cluster is assured of its mutual independency and randomness, the waiting time for the occurrence, or the cluster interval, is distributed as purely exponential. This is one of the remarkable properties of the rain cluster.

In the case study, the rain clusters and no-rain runs are identified from the daily rainfall records from 1975 to 1986 at Hitoyoshi, Kyushu, Japan. The characteristics of the rain cluster and no-rain run as well as the cluster interval are analyzed statistically, and their hitograms are created. Based on these histograms, the probability distributions of these new variables are estimated. It is important to note that these distributions can be regarded as proper or authentic because they are obtained based on the mutually independent random variables. Owing to this authenticity of the probability distributions, many application fields, such as estimation of mean rainfall over a basin, prediction of daily rainfalls, etc., are foreseen and the results are bound to be more reliable than before.

Finally, the author would like to express his sincere gratitude to the officials of Hitoyoshi Weather Station, Japan Meteorological Agency, who kindly showed and explained the activities of the station and gave him the valuable information. Also the auther's foremost thanks go to the officials of Tsuruta Dam Management Office, Ministry of Construction, Japan, and the staff of Sendai-gawa Power Station, Electric Power Development Co., Japan, for their vast cooperation.

REFERENCES

1) W. Feller, An Introduction To Probability Theory And Its Applications, Vol. I, 3rd. ed., Wiley, 1968, p.125.
2) W.J. Conover, Practical Nonparametric Statistics, 2nd. ed., Wiley, 1980, pp.158-169.
3) Op. cit., p.82, pp.216-223.
4) Op. cit., pp.78-80.
5) For the formal definition of the "random variable", see W. Feller, op. cit., pp.212-219.
6) Vujica Yevjevich, Stochastic Processes In Hydrology, Water Resources Publications, 1972.
7) Takeharu Etoh, et. al., On the autocorrelation function of a BRP process, Proc. of JSCE No. 351/II-2 (Hydraulic and Sanitary Eng.), Nov. 1984, pp.137-145. (Japanese)
8) Takeharu Etoh, et.al., A MPP model of daily precipitation series, Proc. of JSCE No. 342, Feb. 1984, pp.171-178. (Japanese)
9) K. Wadachi, et.al., Encyclopedia Of Meteorology, 8th. ed., Tokyo-do Shuppan, 1986, p.350. (Japanese)
10) K. Wadachi, et. al., op.cit., p.223, p.300.
11) W. Feller, op.cit., p.42.
12) Monthly reports, April 1975, issued by Kumamoto Observatory, Miyazaki Observatory, and Kagoshima Observatory. (Japanese)

13) W. Feller, op.cit., pp.146-148.
14) W. Feller, op.cit., pp.458-460.
15) A.M. Mood, et. al., Introduction To The Theory Of Statistics, 3rd ed., McGraw-Hill, 1974, pp.538-543.
16) T.W. Anderson, An Introduction To Multivariate Statistical Analysis, 2nd ed. Wiley, 1984, p.66, p.103.
17) Masashi Nagao, Statistical estimation of theoretical curves between frequency and time distribution ratio of rainfalls, Proc. of JSCE, No. 243, Nov. 1975, pp.33-46. (Japanese)

NOTATIONS

$P\{A\}$: probability of event A.
$P\{B\}$: probability of event B.
$P\{AB\}$: probability of the event "both A and B occur jointly".
i : serial # of the day in the cluster.
$f(x)$: exponential density.
x : independent random variable of the exponential density
α : parameter of the exponential density.
x_o : truncation point of the variable x.

$f_t(x)$: truncated exponential density.

$p(n)$: truncated Poisson distribution.
n : length (in days) of a rain cluster.
ν : a parameter of the truncated Poisson distribution, and also a parameter of the truncated Poisson and gamma compound density.
m_n : mean value of n.

$h(t)$: compound density of truncated Poisson and gamma.
t : amount of rainfall in a rain cluster.
λ : a parameter of the truncated Poisson and gamma compound density.
m_t : mean value of t.

$p(k)$: truncated geometric probability distribution.
k : length (in days) of a no-rain run.
p : a parameter of the truncated geometric probability distribution.
m_k : mean value of k,

m_x : mean value of x