



Analysis and prediction of streamflow and precipitation data
by Alfred Benjamin Cunningham

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE in Civil Engineering
Montana State University
© Copyright by Alfred Benjamin Cunningham (1971)

Abstract:

The double mass analysis and a related method of data synthesis are used to develop a computerized data analysis and generation model. Total monthly volumes of precipitation and streamflow are the types of data for which the model is designed. The double mass analysis is used to check the consistency of the data at a particular station. Then, using the equation of the double mass curve, periods of missing record are synthesized for the station in question. A comparison can be made to determine if cyclic variations exist between the synthetic and actual data. In the event that cyclic variations do occur, correction factors can be applied to improve the accuracy of the synthetic data.

Although the data analysis and generation model was developed for use on the streamflow and precipitation data in Montana, its general structure does not restrict its use to any one geographic region.

Statement of Permission to Copy

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Montana State University, I agree that the Library shall make it freely available for inspection. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by my major professor, or, in his absence, by the Director of Libraries. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature

Albert B. Cunningham

Date

December 3, 1971

ANALYSIS AND PREDICTION OF STREAMFLOW AND PRECIPITATION DATA

by

ALFRED BENJAMIN CUNNINGHAM

A thesis submitted to the Graduate Faculty in partial
fulfillment of the requirements for the degree

of

MASTER OF SCIENCE


in

Civil Engineering

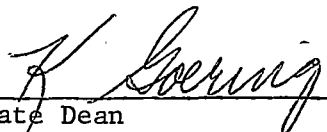
approved:



Head, Major Department



Chairman, Examining Committee



Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

December, 1971

ACKNOWLEDGMENTS

The author wishes to extend his thanks to the faculty of the Civil Engineering and Engineering Mechanics department of Montana State University for their willing assistance, and especially to Professor T. T. Williams for his help and guidance in preparing this thesis.

Thanks are also extended to the author's wife who, along with Mrs. Neta Eckenweiler prepared and typed this thesis.

TABLE OF CONTENTS

	Page Number
VITA.	ii
ACKNOWLEDGEMENTS.	iii
LIST OF FIGURES AND TABLES.	vi
ABSTRACT.	ix
INTRODUCTION	1
STATE WATER PLANNING MODEL.	1
SUMMARY OF PROBLEMS WITH EXISTING DATA.	4
STATEMENT OF OBJECTIVES	4
LITERATURE REVIEW.	6
MODELING TECHNIQUES FOR HYDROLOGIC STUDIES.	6
Mechanical Analog.	6
Electric Analog.	6
Digital Computer	7
STREAMFLOW SYNTHESIS.	7
PRECIPITATION SYNTHESIS	8
CONSISTENCY CHECK FOR STREAMFLOW AND PRECIPITATION DATA	8
DEVELOPMENT OF TECHNIQUE FOR DATA ANALYSIS AND GENERATION:	10
DATA ANALYSIS	10
Consistency Check.	10
Correction of Inconsistent Data.	11

DATA GENERATION.....	15
Criteria Necessary for Synthetic Data Generation	15
Double Mass Extension.....	16
Multiple Regression.....	17
MODEL STRUCTURE.....	22
DATA ANALYSIS SUBROUTINE.....	22
Discontinuity Detection.....	22
Evaluation of the Data Analysis Subroutine.....	26
DATA CORRECTION AND GENERATION SUBROUTINE.....	27
Data Correction.....	27
Data Generation.....	27
Evaluation of the Data Correction and Generation Subroutine.....	28
PRESENTATION OF RESULTS.....	30
DESCRIPTION OF RESULTS.....	30
SIGNIFICANT PARAMETERS.....	30
DISCUSSION OF RESULTS.....	50
ACCURACY VS. CORRELATION.....	50
EFFECTIVENESS OF DOUBLE MASS ANALYSIS.....	52
CYCLIC VARIATIONS.....	54
OPERATION PROCEDURE FOR DATA ANALYSIS AND GENERATION MODEL.....	56

	Page Number
SUMMARY AND RECOMMENDATIONS FOR FURTHER RESEARCH.....	58
APPENDIX A. LITERATURE CONSULTED.....	61
APPENDIX B. PROGRAM LISTING.....	62

LIST OF FIGURES AND TABLES

Figure	Title	Page Number
1	ADJUSTMENT OF INCONSISTENT DOUBLE MASS POINTS. . . .	12
2	DOUBLE MASS EXTENSION.	18
3	MULTIPLE LINEAR REGRESSION	20
4	DETECTION OF DISCONTINUITIES IN DOUBLE MASS ANALYSIS	23
5	GENERATION OF SYNTHETIC DATA BY DOUBLE MASS EXTENSION	29
6-A	DOUBLE MASS DIAGRAM - HYALITE CREEK AT RANGER STATION VS. GALLATIN RIVER AT GALLATIN GATEWAY . .	35
6-B	COMPARISON OF DATA FOR HYALITE CREEK - ACTUAL DATA VS. SYNTHETIC DATA	36
7	PER CENT ERROR VS. CORRELATION COEFFICIENT	37
8-A	DOUBLE MASS DIAGRAM - POWDER RIVER AT LOCATE VS. YELLOWSTONE RIVER AT SIDNEY.	38
8-B	DATA COMPARISON FOR POWDER RIVER - ACTUAL DATA VS. SYNTHETIC DATA	39
9-A	DOUBLE MASS DIAGRAM - GALLATIN RIVER AT GALLATIN GATEWAY VS. MADISON RIVER AT WEST YELLOWSTONE. . .	40
9-B	DATA COMPARISON FOR GALLATIN RIVER - ACTUAL DATA VS. SYNTHETIC AND ADJUSTED SYNTHETIC DATA.	41
10-A	DOUBLE MASS DIAGRAM - YELLOWSTONE RIVER AT YELLOWSTONE LAKE VS. MADISON RIVER AT WEST YELLOWSTONE.	42
10-B	DATA COMPARISON FOR YELLOWSTONE RIVER - ACTUAL DATA VS. SYNTHETIC AND ADJUSTED SYNTHETIC DATA.	43
11-A	DOUBLE MASS DIAGRAM - PRECIPITATION AT WYOLA VS. PRECIPITATION AT SIDNEY.	44

Figure	Title	Page Number
11-B	DATA COMPARISON FOR WYOLA PRECIPITATION - ACTUAL DATA VS. SYNTHETIC DATA.	45
12-A	DOUBLE MASS DIAGRAM - PRECIPITATION AT MONTANA STATE UNIVERSITY VS. PRECIPITATION AT TRIDENT. . .	46
12-B	DATA COMPARISON FOR MSU PRECIPITATION - ACTUAL DATA VS. SYNTHETIC DATA.	47
13-A	DOUBLE MASS DIAGRAM - PRECIPITATION AT PRYOR VS. PRECIPITATION AT BILLINGS.	48
13-B	DATA COMPARISON FOR PRYOR PRECIPITATION - ACTUAL DATA VS. SYNTHETIC DATA.	49

Table	Title	Page Number
I	OCCURRENCES WHICH CAUSE DISCONTINUITIES IN DOUBLE MASS ANALYSIS.	14
II	DATA FOR THE PRECIPITATION AND STREAMFLOW STATIONS WHICH WERE STUDIED	33

ABSTRACT

The double mass analysis and a related method of data synthesis are used to develop a computerized data analysis and generation model. Total monthly volumes of precipitation and streamflow are the types of data for which the model is designed. The double mass analysis is used to check the consistency of the data at a particular station. Then, using the equation of the double mass curve, periods of missing record are synthesized for the station in question. A comparison can be made to determine if cyclic variations exist between the synthetic and actual data. In the event that cyclic variations do occur, correction factors can be applied to improve the accuracy of the synthetic data.

Although the data analysis and generation model was developed for use on the streamflow and precipitation data in Montana, its general structure does not restrict its use to any one geographic region.

Chapter 1

INTRODUCTION

In the last several years, the proper development and management of water resources has become an issue almost everywhere. Previously the greatest concern had been to maintain an adequate water supply for large urban areas; however, it is now becoming apparent that proper development and management is needed everywhere if the optimum use of available water resources is to be achieved. The state of Montana, even with its abundant water resources and sparse population, is no exception. In fact, because so many other areas of the country depend on the water originating in Montana, the need for management in Montana is indeed great.

Until recently, there has not been a concentrated effort to put water resource planning and management techniques to use in this state. However, with the organization of the Montana Water Resources Board out of the old Water Conservation Board in 1967, and the subsequent increase in public support, water resource development and management has become a top priority issue.

STATE WATER PLANNING MODEL

In 1968 the Departments of Civil Engineering & Engineering Mechanics and Industrial & Management Engineering at Montana State University contracted with the Montana Water Resources Board to

develop a statewide water planning model. Although there are several types of modeling techniques in use, the approach taken was to develop a mathematical model for use on the digital computer. This is accomplished by deriving mathematical expressions to represent the interaction of certain hydrologic parameters and programming these expressions for a computer solution. Once these expressions have been derived, it is possible to determine the distribution of the ground and surface water throughout the state for any specified time interval. Having this capability, it will then be possible to evaluate the effects of proposed projects on the state's water resource system.

To accomplish this objective, the state water model will utilize large quantities of hydrologic and geologic data - most of which have never been collected on a regular basis. It is apparent, therefore, that a key to a successful model lies in developing accurate methods of estimating missing data.

Two types of data which are vital to the model are (1) records of the precipitation which falls on the state and (2) records of the streamflow which occurs within the state. Although these data are collected for Montana by the National Weather Service and the U.S. Geological Survey, the records at many locations are presently inadequate. This is primarily because precipitation and streamflow records in Montana range in length from one or two years up to fifty or more years. Also, the data for some locations have not been

continuously collected throughout the lifetime of the station. In fact an examination revealed that missing record periods are present in approximately 80% of the available precipitation and streamflow records for Montana. Thus if a method could be developed whereby periods of missing record could be synthesized accurately, a significant improvement to the State Water Planning Model would be made.

Another inadequacy in the data that must be recognized is that some periods of record are "inconsistent." It is important here to realize what is meant and not confuse the term "erroneous" with the term "inconsistent." The record for a particular station is "erroneous" if the gage inaccurately measures the true quantity of precipitation falling or streamflow passing. The record is "inconsistent" if some natural or man-caused occurrence (such as the physical relocation of a gage to a different site) causes a shift or discontinuity in the data.

For example, consider a stream gaging station at which data have been collected for a long period of years. If this station were then to be moved upstream past a major tributary to the river, the streamflow record after the move would be inconsistent with the record before the move. That is, the recorded flow would then be less compared to what it would have been if the gage had not been moved. Notice in this example that the gaging station is assumed to record the true flows at each location. But if the record from this station is assumed to consist of the two segments of data collected from the

two sites without adjusting the data from one, then the record would be said to be "erroneous."

It is easily seen that if the presence of inconsistent records is not detected, any attempt to use these records in a watershed model or to synthesize missing data could result in serious error. Therefore a necessary prerequisite for data synthesis is that existing data which is to be used, first be tested for consistency.

SUMMARY OF PROBLEMS WITH EXISTING DATA

The problems associated with the streamflow and precipitation data for the State Water Planning Model can be summarized in the following way.

The primary problem with the data is that periods of data are missing from about 80% of the existing records. Therefore a way must be found to generate synthetic data in order to have complete records at all desired locations.

The second problem which is less significant (though still important) is that some existing records are inconsistent and must be corrected before being used for data generation or any other purpose. Therefore a way must be found to test existing streamflow and precipitation data for consistency.

STATEMENT OF OBJECTIVES

Based on the foregoing statements concerning the need for the development of a "Data Analysis and Generation Model," the following objectives are stated.

- (1). Construct a computerized "Data Analysis Model" which will check the consistency of existing precipitation and streamflow data.
- (2). Construct a computerized "Data Generation and Correction Model" capable of synthesizing missing periods of monthly streamflow and precipitation volumes - as well as correcting existing data which are inconsistent.
- (5). Describe the operation procedures for these models which will yield the best possible results.

Chapter 2

LITERATURE REVIEW

A survey of pertinent literature was conducted to analyze the various hydrologic modeling techniques as well as the various methods of data analysis and generation.

MODELING TECHNIQUES FOR HYDROLOGIC STUDIES

Three types of modeling techniques which are currently being used in hydrologic studies are given by Mount, (1965). These three types are (1) the mechanical-analog model, (2) the electric-analog model, and (3) the digital computer model.

Mechanical Analog

An example of applying a mechanical-analog model to a hydrologic study would be to use a stretched membrane to represent the effect of pumping water from a well. As the well is pumped, the water table assumes the shape of an inverted cone with the apex at the well. The same shape occurs in a membrane when it is pressed with a sharp instrument. In this case the elastic properties of the membrane are analogous to the water transmitting properties of the aquifer and the magnitude of the point force applied to the membrane is analogous to the rate of withdrawal from the aquifer.

Electric Analog

Electric-analog models, however, appear to have greater utility than mechanical-analog models because electric properties have

analogues in many physical systems. However, in hydrologic studies, electrical systems can be used to model certain aspects of the hydrologic cycle. For example, an electric-analog model could be constructed to depict the changes in ground and surface water conditions for a given area. Here, the ground and surface storage would be represented by capacitance streamflow, precipitation, and ground water movement would be represented by current; and soil transmissibility would be represented by resistance.

Digital Computer

Although the electric-analog and the mechanical analog models are useful in many cases, the digital computer model is more versatile and therefore used more often. The application of digital computer modeling to hydrology involves determining mathematical relations which represent the interaction of hydrologic parameters. For example, consider the Stanford Watershed Model which is described by Linsley and Crawford, (1966). In this model virtually every aspect of the hydrologic cycle has been represented mathematically and programmed for a computer solution. Once the required hydrologic data have been obtained for a particular watershed, the Stanford Watershed Model can be used to predict the outflow hydrograph resulting from any theoretical storm which could occur on the watershed.

STREAMFLOW SYNTHESIS

A computer solution for estimating periods of missing streamflow record was developed by Beard, (1967). The procedure was to

make the existing data at a particular gaging station a function of the data from a group of surrounding stations. The method of least squares is used as the basis for this analysis and results in the development of a linear relation between the data for the particular station and the data for the surrounding stations. Once this equation is determined, Beard's program can then fill in any periods of missing record for the particular station.

PRECIPITATION SYNTHESIS

Linsley, Kohler, Paulhus, (1950) presented a method which graphically determines the distribution of rainfall from a particular storm over a given area. This is done by first plotting the recorded rainfall at each of the gaging stations for the area in question. Then, isohyets or "rainfall contours" are drawn to indicate how the rainfall was distributed over the area. Thus the rainfall could be estimated for any station which was not operating at the time of the storm. Though it would be quite time consuming, an entire period of record at a given station could be estimated by repeating this procedure for each storm that occurred during the time the gage was not operating.

CONSISTENCY CHECK FOR STREAMFLOW AND PRECIPITATION DATA

A survey of available literature revealed that the standard method for testing the consistency of hydrologic data is the "double mass analysis" (DMA). In fact no other method was found in the literature. A development of the DMA for use in testing the

consistency of precipitation data is presented by Linsley, Kohler, & Paulhus, (1958). Although the DMA is primarily used for precipitation data, it is pointed by Linsley, Kohler, Paulhus, (1949), that the DMA can be used to test the consistency of streamflow data as well.

It is because of this capability for testing the consistency of both types of data that the double mass analysis was chosen as the basis for the data analysis and generation model.

R. Singh (1968) developed a computer solution for the double mass analysis. However, his program was deemed inappropriate for this study because it could not detect the presence of more than one period of inconsistent data.

Chapter 3

DEVELOPMENT OF TECHNIQUES FOR DATA ANALYSIS AND GENERATION

Based on the results of the literature survey, techniques were developed for both the analysis of existing data and generation of synthetic data.

DATA ANALYSIS

As previously stated, the double mass analysis is used by hydrologists to test the consistency of both precipitation and stream-flow data. Consider now how the DMA is applied to precipitation data.

Consistency Check

The consistency of precipitation data is checked by comparing the accumulated precipitation at a given station with the accumulated precipitation for a group of surrounding stations. The procedure is to plot the accumulated rainfall values at the particular station as the dependent variable and the concurrent accumulated rainfall for the surrounding stations as the independent variable. If the records from the stations are well correlated, as they should be if the stations are in the same hydrologic unit, the points should plot approximately on a straight line. However if a slope change is necessary to fit line segments through the points, it is highly probable that an inconsistency exists in the data for the dependent station. If a slope change is detected, it is important to check the history of the dependent station to confirm the existence of the

data inconsistency. If the gage history reveals the occurrence of an event which could cause a slope change in the data, then it is almost certain that an inconsistency in the dependent station record exists. If the slope change cannot be substantiated from the gage history, it is a judgment decision as to whether the slope change does in fact indicate a discontinuity in the data. Care must be taken in this case not to interpret the natural scatter of the points as an inconsistency in the data. The possible reasons for discontinuities in both precipitation and streamflow data are discussed later in this chapter.

The above discussion is an explanation of the procedure involved in performing the double mass analysis. Although this procedure was discussed with respect to precipitation data, it is important to note that streamflow data can be analyzed in exactly the same manner. The only difference is that these two types of data will have different units. The streamflow data used in this study will have the units of acre-ft/month while the precipitation data will be in inches/month.

Correction of Inconsistent Data

Once the DMA has detected a discontinuity in the record of the dependent station, correction of the erroneous data is easily done. The procedure is to multiply each erroneous double mass ordinate by the ratio of the slopes of the two line segments constructed through the double mass points.

To illustrate, a double mass curve is shown in Figure 1. The individual data values are shown along with the resulting double

<u>y</u>	<u>Y</u>	<u>x</u>	<u>X</u>
1.0	1.0	2.0	2.0
2.5	3.5	5.0	7.0
1.0	4.0	2.0	9.0
.5	6.0	1.0	10.0
2.0	6.5	4.0	14.0
.5	9.5	1.0	15.0
3.0	10.5	1.5	16.5
1.0	12.5	.5	17.0
2.0	14.5	1.0	18.0
2.0	16.5	1.0	19.0

y = Dependent station data. Y = Ordinates on double mass diagram.
 x = Surrounding station data. X = Abscissa values for double mass diagram.

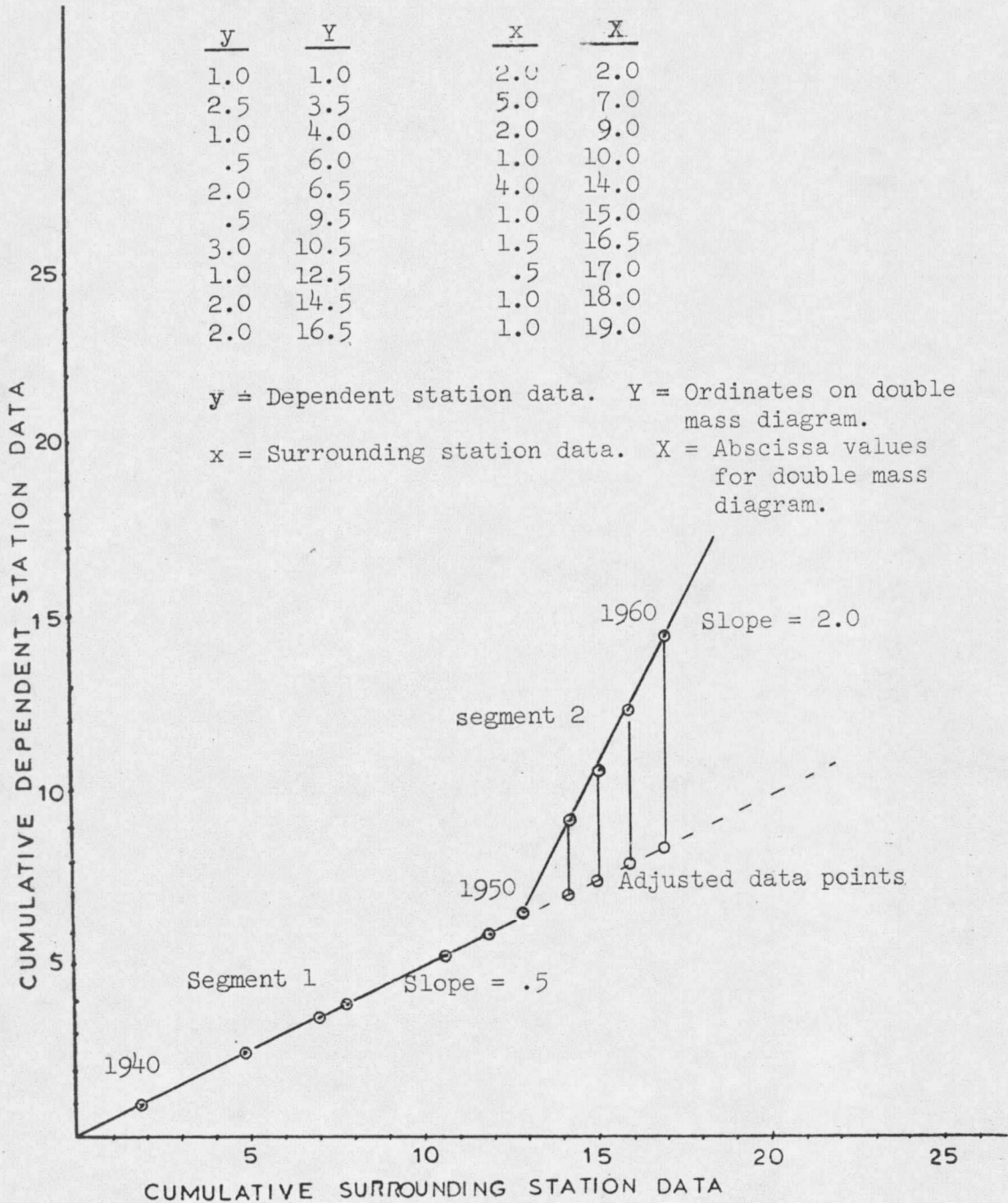


FIGURE 1. ADJUSTMENT OF INCONSISTENT DOUBLE MASS POINTS

mass coordinates to demonstrate how the double mass lines were constructed. Plotting the double mass coordinates (Y vs. X), shows that a slope change occurs approximately in 1950. Assuming that the slope change does in fact represent a discontinuity in the dependent station record, the record from 1950 to 1960 can be corrected by multiplying the segment of the ordinates which lie above the discontinuity by the ratio $.5/2.0$ so as to be consistent with the points on segment 1.

Thus far it has been assumed that if a discontinuity is found by a double mass analysis the record for the dependent station is automatically in error. However, in the event that only one surrounding station is used as the independent station, it is equally possible that a discontinuity could be due to erroneous data for the independent station. In this case the record for both the independent station and the dependent station must be examined if a slope change is found by the double mass analysis. The most common causes for the occurrence of discontinuities in streamflow and precipitation data are listed in Table I. Where precipitation data is concerned, it has been observed by the National Weather Service, according to Linsley, Kohler, and Paulhus (1958), that a change in gage location can be a significant factor in causing the data to be inconsistent. In some cases a change in location of less than five miles can cause significant error in the gage record. Changes in the exposure of a gage can also be a significant factor in affecting the accuracy of the gage record. This happens in areas where trees or vegetation are allowed to grow up around a gage, thus affecting the catch.

TABLE I
OCCURENCES WHICH CAUSE DISCONTINUITIES
IN THE DOUBLE MASS ANALYSIS

PRECIPITATION DATA

1. Changes in gage location
2. Changes in the gage exposure
3. Changes in instrumentation
4. Changes in observation techniques

STREAMFLOW DATA

1. Construction of hydraulic structures
2. Changes in diversion practices
3. Changes in gage location
4. Changes in observation procedure
5. Changes in instrumentation
6. Erosion or sedimentation in vicinity of gage

Variations in observation procedure or changes in instrumentation can also affect the accuracy of raingage data. For example changing from a simple volumetric recorder to a continuous recorder could change the accuracy of a monthly gage record considerably. Also if the observation procedure for a simple volumetric recorder is

changed so that the gage is checked every month instead of daily, a decrease in the accuracy of the monthly precipitation values could be observed. It is these types of "accuracy changes" which can also cause inconsistencies in rain gage data.

In the case of monthly streamflow data, changes in instrumentation gage location and observation procedure could affect the accuracy of the record in much the same way as they affect precipitation records. However other possible causes for discontinuities in streamflow data are changes in diversion practices and construction of hydraulic structures. The occurrence of either event above a stream gaging station will almost certainly cause a discontinuity in the record.

DATA GENERATION

Once all existing periods of data for a particular station have been checked for consistency, the next task is to fill in the missing record periods for which there are concurrent records from surrounding stations. Before the various alternatives are explained it is necessary to define the criteria needed to permit missing data to be synthesized. These criteria are presented below.

Criteria Necessary For Synthetic Data Generation

- (1) The data used to obtain a relationship between the dependent variable (the data for the dependent station) and the independent variable (the data for the surrounding stations) must be tested and found to be consistent.

- (2) The independent variable data which is used to fill in missing periods of dependent variable data, must be consistent.

In the light of these criteria two alternate methods for data generation were explored.

Double Mass Extension

The first method considered was an extension of the double mass curve. The procedure is to substitute known "X" (double mass dependent variable values) values for the period of missing record, into the established double mass curve equation of the form:

$$Y = MX + b$$

where X = The independent variable.

Y = The dependent variable.

M = The slope of double mass line.

b = The intercept of double mass line.

The result is the generation of synthetic values of "Y" for the missing record period. The points on the double mass curve are determined by the relationship:

$$Y_n = \sum_{i=1}^n y_i$$

where Y_n = The ordinate of the n^{th} double mass point.

and

y_i = The individual base station data values.

$$X_n = \sum_{i=1}^n x_i$$

X_n = The abscissa of the n^{th} double mass point.

x_i = The individual surrounding station data values.

The actual synthetic missing data values are found by subtracting each "Y" from the "Y" value immediately succeeding it.

That is:

$$y_i = Y_i - Y_{i-1}$$

y_i = The i th synthetic data value.

Y_i and $i-1$ = The synthetic double mass ordinates used to obtain y_i .

Thus the necessary missing record for the dependent station is generated in this manner.

Multiple Regression

The second method for data generation which was explored was the method of "multiple regression," which is a direct application of the Theory of Least Squares.

Multiple regression in this case would involve using a period of known record to make the dependent station data a function of the data from surrounding stations. This is accomplished by fitting the following type of polynomial to the existing data.

$$y = C_0 + C_1 x_1 + C_2 x_2 \dots C_n x_n$$

where

y = data for dependent station

$x_{1,n}$ = Concurrent data for surrounding stations.

$C_{0,n}$ = coefficients which are determined by the least square fit of the polynomial to the data.

The least squares fit of the polynomial to the data results

in the determination of the coefficients C_0, C_1, \dots, C_n . Once these are known, periods of missing dependent station record (y values) can be determined merely by knowing the concurrent "x" values for the surrounding stations.

At first glance little similarity is apparent between the two methods for data generation. However, further examination shows that there is in fact, a great deal of theoretical similarity. It can be shown for a special case that theoretically double mass extension and multiple regression will produce the same synthetic data values. This special case occurs when the data at the dependent station is made a function of the data from only one neighboring station. The proof is as follows:

Consider first how synthetic data are generated using an extension of the line of best fit through the double mass points.

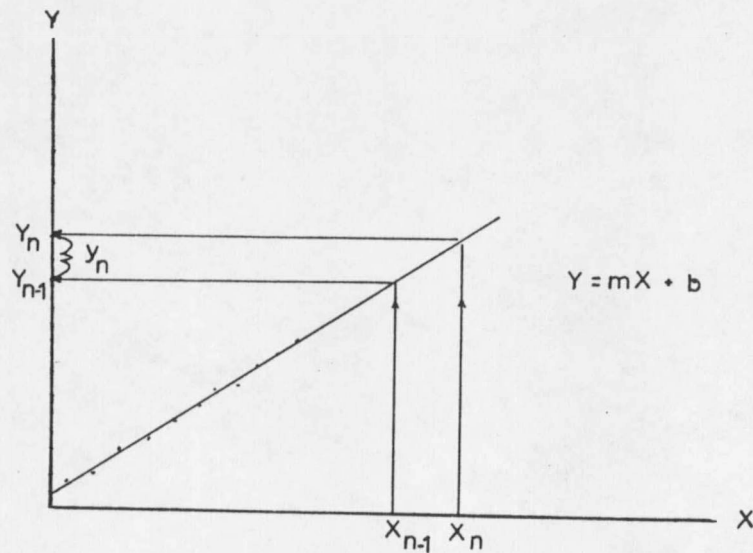


Figure 2. DOUBLE MASS EXTENSION

In the previous diagram,

$$Y_n = y_i \quad (i = 1, \dots, n)$$

$$X_n = x_i \quad (i = 1, \dots, n)$$

y_i = The individual data values for the dependent station

x_i = The individual data values for the surrounding stations

To estimate any synthetic data value y_n , the following procedure is used.

$$y_n = Y_n - Y_{n-1} = (mX_n + b) - (mX_{n-1} + b)$$

This can be rewritten as:

$$y_n = m(X_n - X_{n-1}) + b - b$$

Since $(X_n - X_{n-1}) = x_n$

$$y_n = mx_n \quad \text{where } m = \text{slope} = \frac{y_n}{x_n} \quad (1)$$

Now consider how synthetic data are calculated from multiple linear regression.

Using concurrent data from the dependent station and the surrounding station, a linear relation is obtained.

$$y_n = c_0 + c_1x_n \quad \text{where } y_n = \text{Individual data values for the dependent sta.}$$

x_n = Individual data values for the surrounding sta.

This relation is represented graphically as follows:

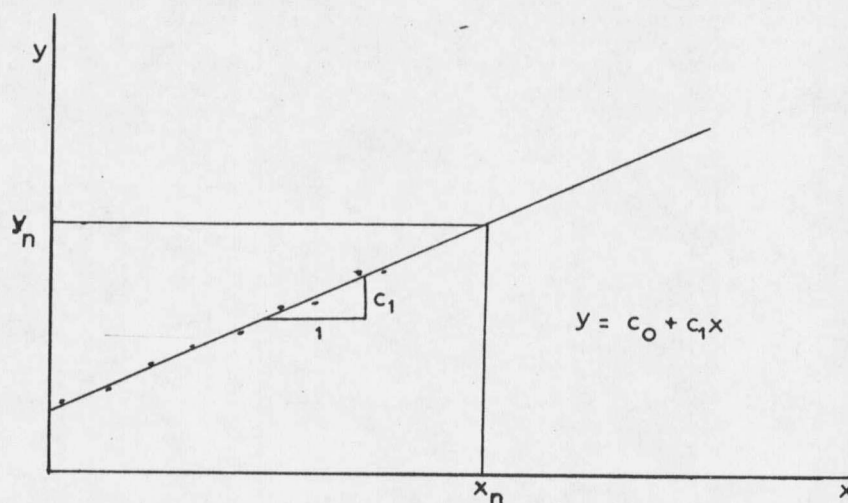


Figure 3. MULTIPLE LINEAR REGRESSION

To obtain an estimate of the synthetic data value y_n the only information needed is the known value for x_n .

Analysis of the above diagram leads to the following equation.

$$\text{since } c = \text{slope} = \frac{y_n - c_0}{x_n - 0}$$

$$\therefore y_n = \frac{(y_n - c_0)}{x_n} x_n + c_0 \quad (2)$$

But E . 1 can also be written: $y_n = \frac{(y_n + c_0 - c_0)}{x_n} x_n$

$$\text{Or: } y_n = \frac{(y_n - c_0)}{x_n} x_n + c_0 \quad (3)$$

Since Eq. 2 and the final form of Eq. 1 are identical, it has therefore been shown that the synthesis techniques of "double mass extension" and "multiple linear regression" will produce the same synthetic data values - provided that the dependent station is made a function of only one surrounding station.

This proof provided the basis for the decision to use double mass extension for the synthetic data generation portion of the model. This decision allowed the computer programming to be done such that information from the data analysis section could be very easily transferred to the data generation section. Had the multiple linear regression technique been used for data synthesis, a more complicated program structure would have been required.

Chapter 4

MODEL STRUCTURE

The data analysis and generation model consists of two distinct units with the output from the first unit (the data analysis subroutine) serving as partial input to the second unit (the data correction and generation subroutine).

DATA ANALYSIS SUBROUTINE

Since the primary purpose of the "Data Analysis Model" is to detect discontinuities in existing data, a method had to be developed whereby the computer would accurately detect both the presence and the location of these discontinuities.

Discontinuity Detection

The method used is presented below. (Refer to Figure 4)

1. First the coordinates of the points on the double mass curve are calculated by a subroutine of the program.

These points would be similar to those plotted in Figure 4.

2. Using the following least squares equation

$$Y = \bar{Y} + M (X - \bar{X}) \quad \text{where } M = \frac{\sum_{i=1}^N (X_i Y_i) - N \bar{X} \bar{Y}}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

the model next calculates the line of best fit through all of the data points (one point at a time). That is the first line of best fit is passed through the first double mass point only, the second line is passed through the

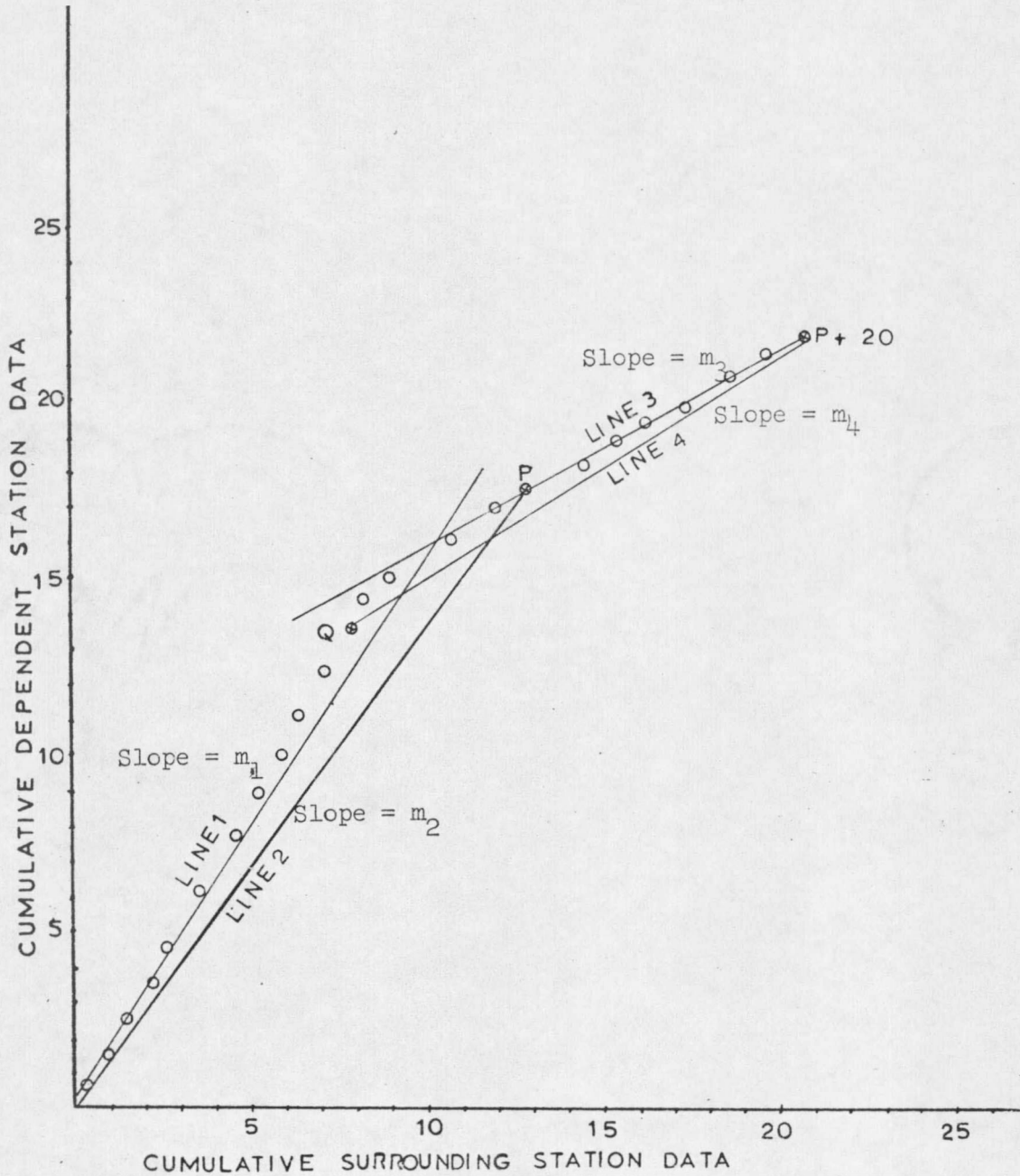


FIGURE 4. DETECTION OF DISCONTINUITIES IN DOUBLE MASS ANALYSIS

first two points etc. The most important parameter obtained from these calculations is the slope "M₁" of the line of best fit. This parameter is used in determining if a discontinuity exists in the data.

3. Along with obtaining the slope "M₁", the model simultaneously calculates the slope of another straight line. This line passes through the particular double mass point and the origin. For example, when the model fits a line through the first four double mass points, it also calculates the equation of the line passing through point #4 and the origin. The slope of this two point line is labeled "M₂".
4. After each calculation of the slope values "M₁" & "M₂", the program checks for a discontinuity by calculating the ratio.

$$R_1 = \frac{M_1 - M_2}{M_1}$$

If the absolute value of "R₁" is greater than the prescribed tolerance, then a discontinuity is assumed. In Figure 4, line 1 represents the line of best fit through all points up to point "P." Line 2 represents the line passing through point "P" and the origin. The slopes of lines 1 and 2 are "M₁" & "M₂" respectively. Assuming that the value of "R₁" exceeds the tolerance, the

presence of a discontinuity has now been established. This also establishes that point "P" lies "above" the discontinuity.

Discontinuity Location

5. Once the presence of a discontinuity is established, the program next attempts to accurately determine its location. To do this the program first fits a line through the first twenty points above point "P". The slope of this line is defined as "M₃." Now another "two point" line is calculated through point "P + 20" and the points lying below point "P" - one point at a time. The slope of this line is labeled "M₄." This process continues until the ratio:

$$R_2 = \frac{M_4 - M_3}{M_3} \quad \text{is exceeded. It must be noted here}$$

that the ratios R₁ & R₂ must be set by the programmer prior to the running the program. It was found that values of R₁ & R₂ in the range .1 - .5 were the most satisfactory, taking into account that R₁ & 2 should be chosen proportionally with the "scatter" of the double mass points. In Figure 4, line 3 represents the line of best fit through all points between point "P" and point "P + 20" - the slope of which is "M₃." Point "Q" is the point at which "R" is assumed to exceed the tolerance.

6. The point of discontinuity is assumed to lie half way between point "P" and point "Q."

$$\text{Break point} = \frac{P + Q}{2}$$

7. Once a point of discontinuity is located, the program now neglects all points below the "break point," thus treating the break point as the new origin of the double mass curve. Starting from this new origin, the model now repeats steps 1 - 6 to check for additional discontinuities.

Evaluation of the Data Analysis Subroutine

The method used in constructing the data analysis model was chosen because of the following advantages:

1. The method allows the model to detect more than one discontinuity.
2. The method lends itself easily to a computer solution.
3. The logic behind this method is simple and straight forward.

This method of data analysis also has certain disadvantages.

These include:

1. The accuracy with which the program locates slope changes is proportional to the "degree of scatter" of the double mass points. This scatter is measured by computing the standard error of estimate of the points about the line of best fit.

2. The tolerance which determines when a break point occurs must be set by the programmer.

It is important to keep these disadvantages in mind, especially when interpreting the results of the Data Analysis Model.

DATA GENERATION AND CORRECTION SUBROUTINE

Keeping in mind the methodology presented in Chapter 2, the program structure is as follows:

Data Correction

1. If discontinuities are detected in the data by the Data Analysis Model, the first step of this program is to correct the existing inconsistent data. To accomplish this, the programmer must first decide how the data is to be corrected based on examination of existing records. Once this has been determined, the proper correction equations are placed in a special subroutine and the data is corrected.

Data Generation

2. Having corrected all faulty data, periods of missing records are now synthesized by the data generation section of the model. The procedure used to accomplish this was simply to program the methodology outlined in Chapter 2. Once this was done, the input needed for data generation were (1) surrounding station data for the periods of missing base station record, and (2) the equation of the line of best fit for the correct period of data on the double mass curve.

As an example of how the data correction and generation model works, consider Figure 5.

Assuming the data on segment I to be in error, the programmer must first decide how the data is to be corrected. That is, it must be decided to move the double mass points on segment I either vertically, horizontally or in both directions. As stated before, this is a judgment decision based on examination of the records for the stations involved. Once this has been decided the points on segment I are corrected in the proper manner to lie on segment II.

Now assume that the known period of base station record ends at point A. Assume also that the equation of line segments II & III is the correct model to be used for data generation. The known abscissa values are now substituted into the correct equation to yield the ordinate value for the missing period. These ordinates are then subtracted in the proper manner to produce the missing data values.

Evaluation of Data Correction and Generation Subroutine

The advantages which justify the use of this approach to data generation and correction are:

1. The exact method of data correction is in each case left to the judgment of the programmer.
2. The model provides for the correction and generation of more than one period of data.

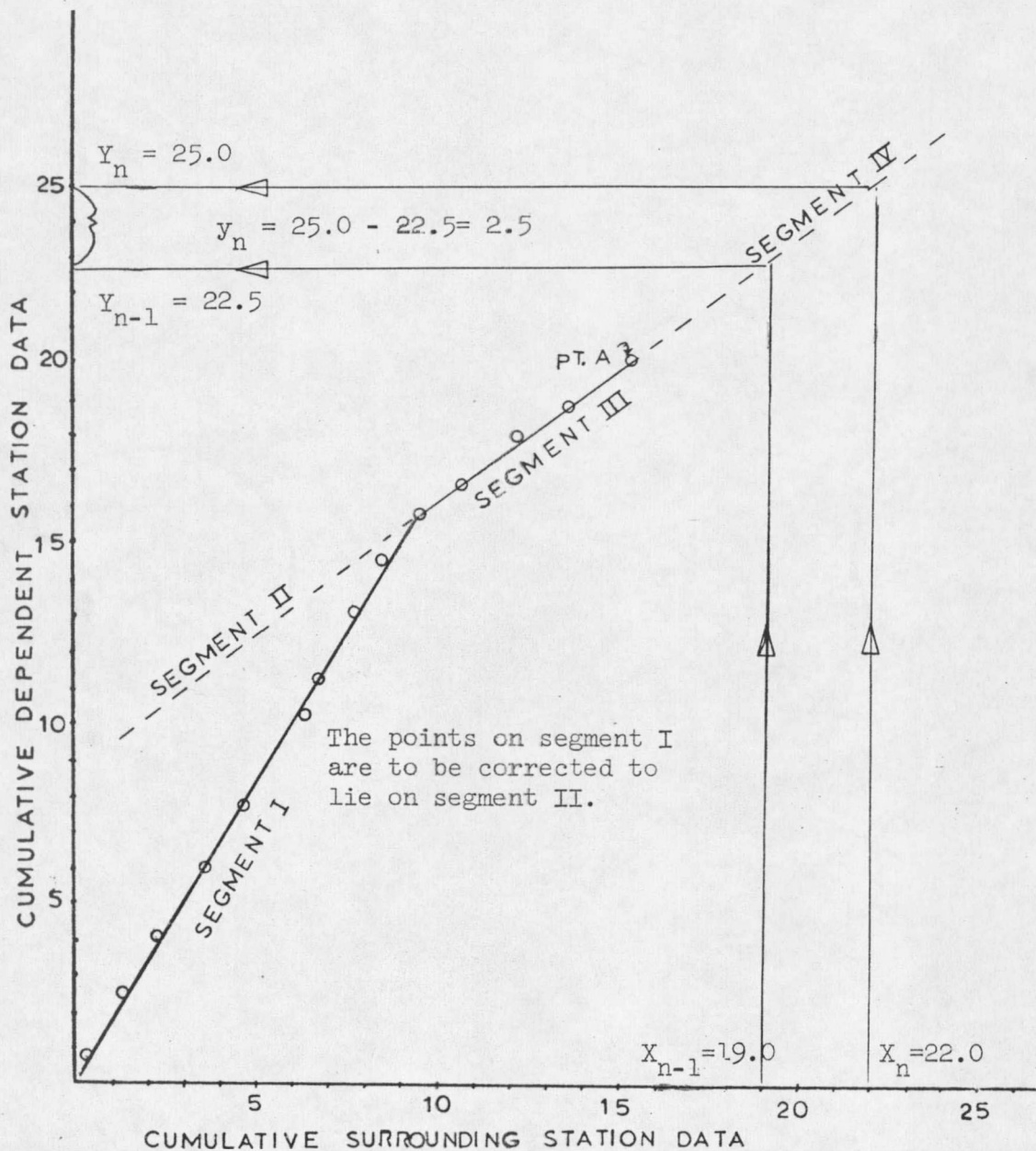


FIGURE 5. GENERATION OF SYNTHETIC DATA BY DOUBLE MASS EXTENSION

Chapter 5

PRESENTATION OF RESULTS

The effectiveness of the data analysis and generation model was tested using data from various gaging stations in Montana. The test procedure was first to check the consistency of the data, and then after making any necessary corrections, to synthesize a period of data. For the purpose of testing, periods of record were synthesized which in fact were already known. Thus a comparison could be made to determine the effectiveness of the model.

DESCRIPTION OF RESULTS

The results which follow consist of double mass curves (which check the consistency of a particular set of data) and graphs of the generated synthetic data plotted with the actual data. Both streamflow and precipitation data are represented in the results. The streamflow data tested came primarily from the Gallatin, Madison, and Yellowstone River drainages, while the precipitation data came from both Gallatin County and Southeastern Montana.

SIGNIFICANT PARAMETERS

Significant parameters which are used in the discussion of the results are the following:

1. Correlation Coefficient
defined as:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n S_x S_y}$$

where

x_i = known base station data

\bar{x} = mean of x data

S_x = STD. deviation of x data

y_i = known surrounding station data

\bar{y} = mean of y data

S_y = standard deviation of y data

n = total number of data points.

2. Per Cent Error defined as:

$$E = \frac{|M_s - M_a|}{M_a} \times 100$$

where

M_s = mean of the generated synthetic data

M_a = mean of the actual data for the period of synthetic data.

Per Cent Error is used to estimate the accuracy with which the synthetic data for a given period compares with the actual data for that same period.

3. Monthly Per Cent Error defined as:

$$e(i) = \frac{M_s(i) - M_a(i)}{M_a(i)} \times 100 \quad (i = 1, \dots, 12)$$

where

$M_s(i)$ = mean of the synthetic data for a given month of the year.

-32-

$M_a(i)$ = mean of the actual data
for a given month of the
year.

$e(i)$ = Average monthly deviations.

TABLE II

DATA FOR THE PRECIPITATION AND STREAMFLOW
STATIONS WHICH WERE STUDIED

STREAMFLOW STATIONS

1. Hyalite Creek (U.S.G.S. # 6-0500)

LOCATION ---At Hyalite Ranger Station 7.3 miles south
of Bozeman, Montana.

DRAINAGE AREA ---48.2 sq. mi.

REMARKS ---Records fair. Flow regulated by Middle
Creek Reservoir since 1951.

2. Gallatin River near Gallatin Gateway (U.S.G.S. #60435)

LOCATION ---7.3 miles south of Gallatin Gateway, Montana.

DRAINAGE AREA ---825 sq. mi.

REMARKS ---Records good. Diversions for about 1400
acres above station.

3. Madison River near West Yellowstone Montana (U.S.G.S. #6-0375)

LOCATION ---1.6 miles east of West Yellowstone, Montana.

DRAINAGE AREA ---420 sq. mi.

REMARKS ---Records good. No diversion or regulation
above gage.

4. Yellowstone River at Yellowstone Lake outlet (U.S.G.S. #6-1865)

LOCATION ---.2 miles downstream from outlet of
Yellowstone Lake.

DRAINAGE AREA --- 1006 sq mi.

REMARKS --- Records good except for winter months which are bad.

PRECIPITATION STATIONS

1. Wyola

LOCATION ---On Interstate 90 fourteen miles from
Montana-Wyoming border.

2. Sidney

LOCATION ---On the Yellowstone River seven miles
from Montana-North Dakota border.

3. Bozeman

LOCATION ---On campus of Montana State University

4. Trident

LOCATION ---Thirty one miles west of Bozeman, Montana.

5. Pryor

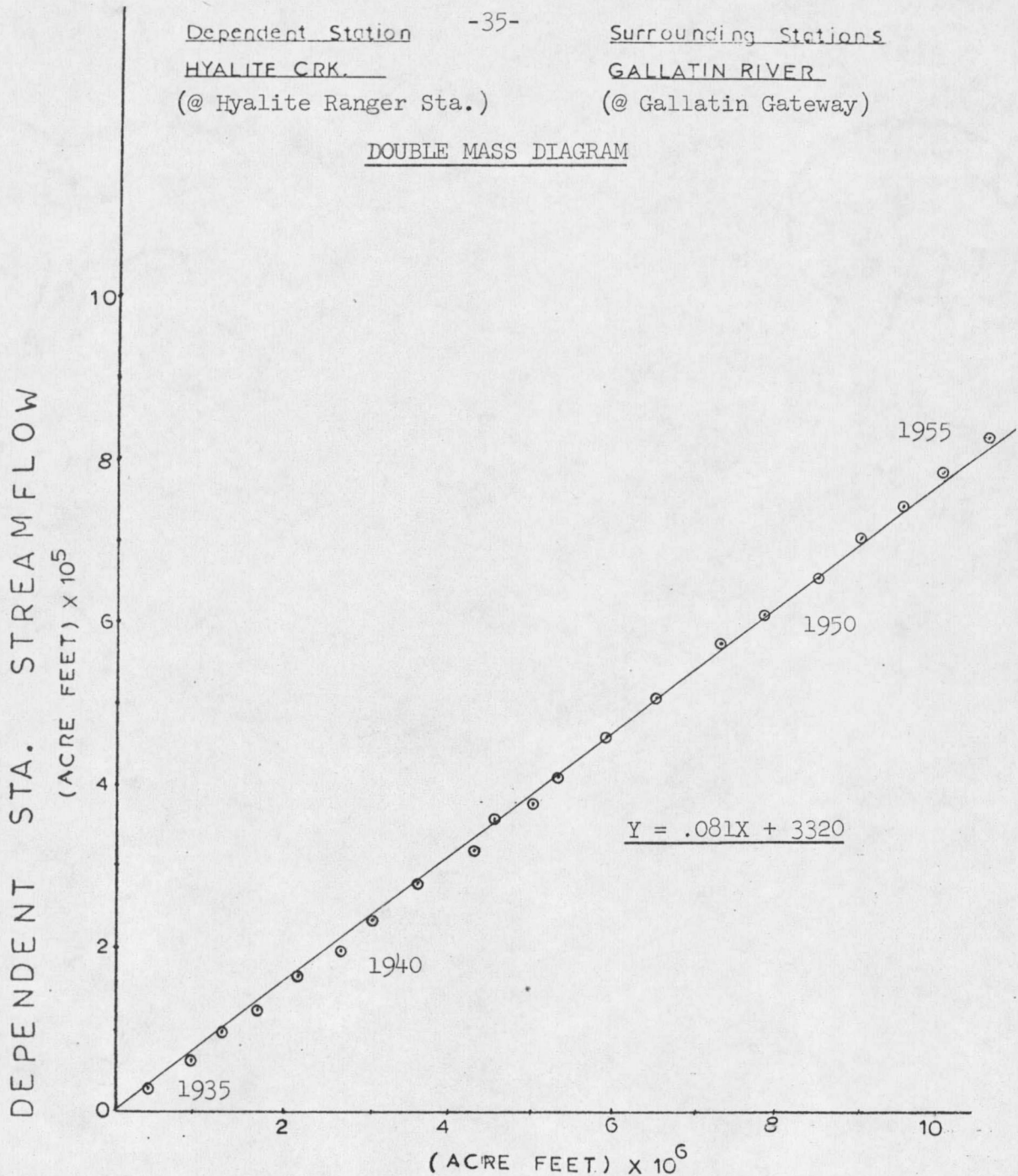
LOCATION ---Thirty seven miles south of
Billings, Montana.

6. Billings

LOCATION. ---Downtown Billings, Montana.

Dependent Station Surrounding Stations
HYALITE CRK. GALLATIN RIVER
(@ Hyalite Ranger Sta.) (@ Gallatin Gateway)

DOUBLE MASS DIAGRAM



SURROUNDING STA. STREAMFLOW

FIGURE 6-A. DOUBLE MASS DIAGRAM :: STREAMFLOW

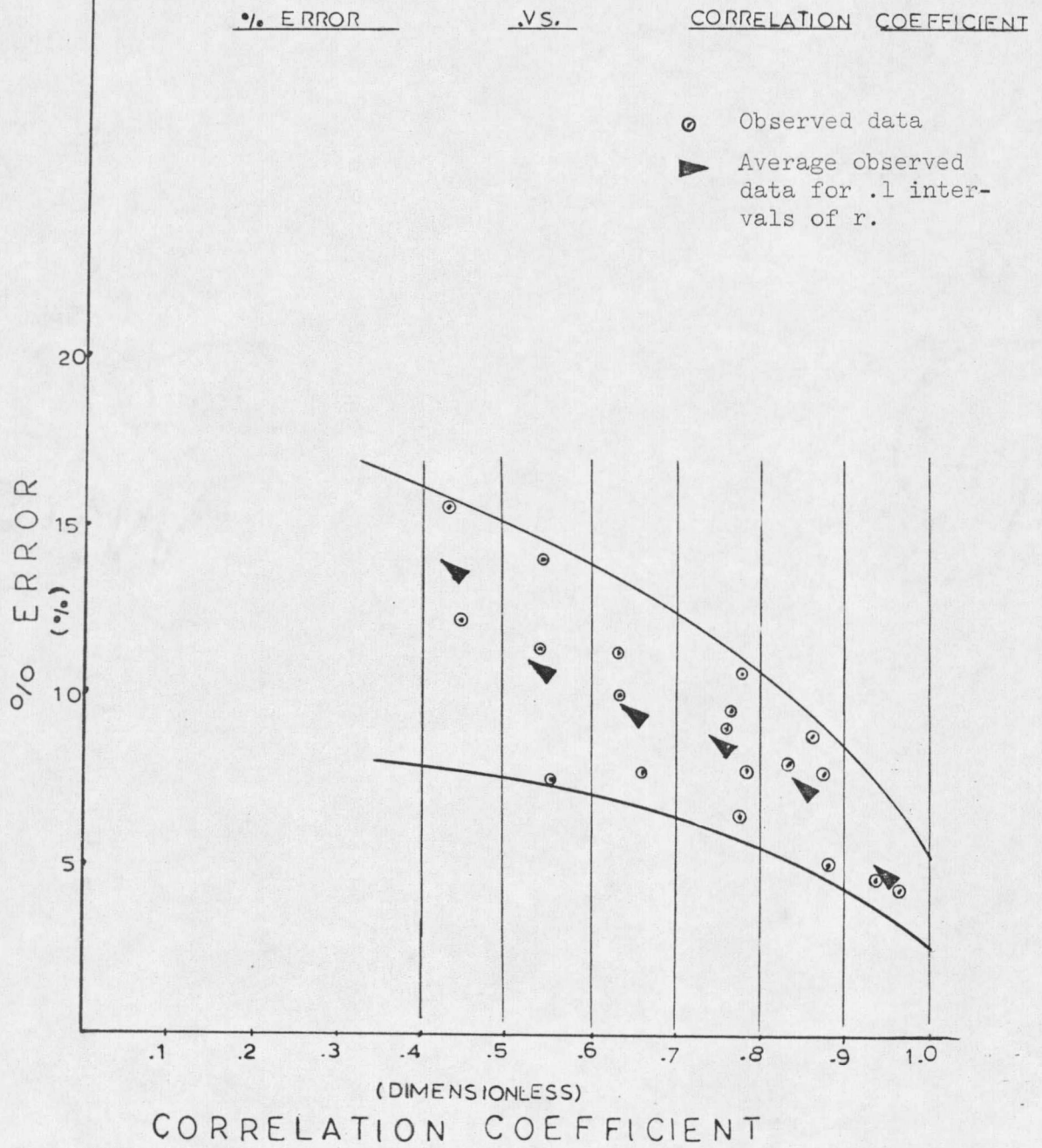


FIGURE 7. PER CENT ERROR vs. CORRELATION COEFFICIENT

Dependent Station

Surrounding Stations

POWDER RIVER

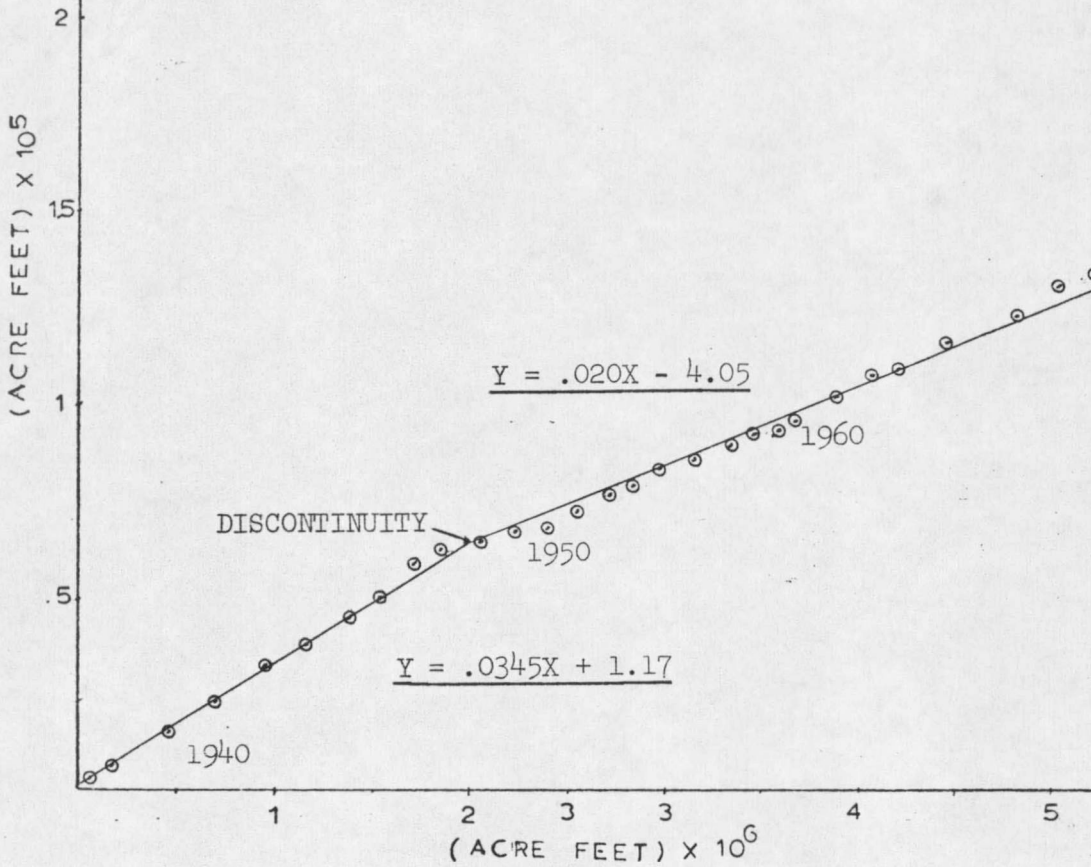
YELLOWSTONE RIVER

(@ Locate)

(@ Sidney)

DOUBLE MASS DIAGRAM

DEPENDENT STA. STREAMFLOW
(ACRE FEET) X 10⁵



SURROUNDING STA. STREAMFLOW

FIGURE 8-A. DOUBLE MASS DIAGRAM: STREAMFLOW

MONTHLY STREAMFLOW
 $\times 10^3$ (ACRE FEET)

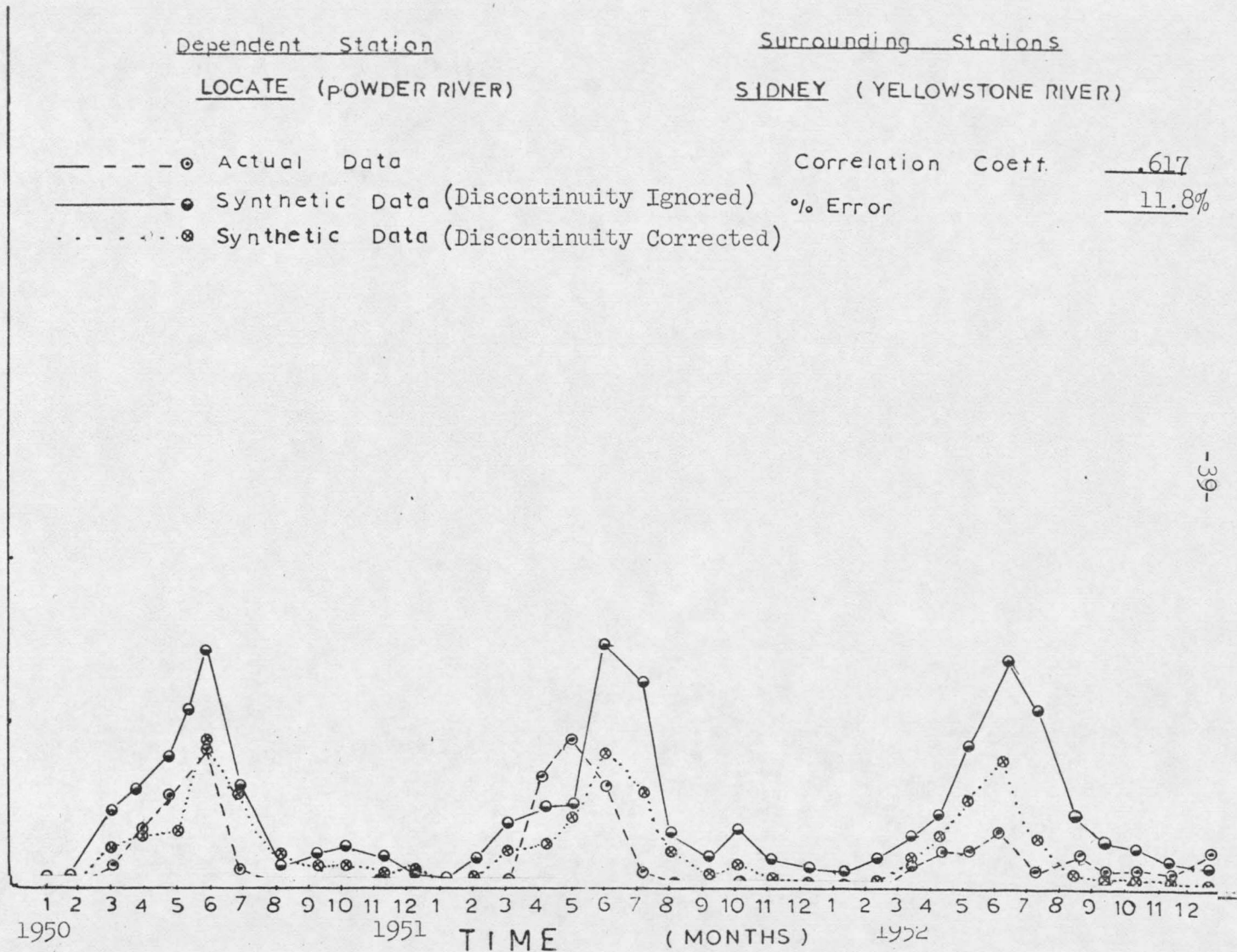


FIGURE 8-B. COMPARISON OF DATA FOR POWDER RIVER

