RAINFALL AND ACIDITY DATA SUMMARY FOR SELECTED UAPSP MONITORING SITES

PAU-CHU PAM CHEN

TECHNICAL REPORT NO. 80
AUGUST 1985

SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

Rand
HEALTH SCIENCES PROGRAM
RAND CORPORATION

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
RAINFALL AND ACIDITY DATA SUMMARY FOR SELECTED UAPSP MONITORING SITES

by

PAU-CHU PAM CHEN
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 80
AUGUST 1985

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER A GRANT TO SIMS FROM
ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Rainfall and Acidity Data Summary for
Selected UAPSP Monitoring Sites

Pao-Chu Pam Chen
Stanford University

1. Introduction.

The Utility Acid Precipitation Study Program (UAPSP) network incorporated an existing six-station network that had been part of EPRI's Sulfate Regional Experiment (SURE). As a result, 5 years of data from these 6 sampling locations are now available, together with 2 years of data for 19 locations locations in the eastern United States (Figure 1). Their primary goal is to operate a high-quality precipitation chemistry network in order to establish temporal and spatial variability and trends with specified accuracy and precision.

This report is to present the rainfall and acidity data summary for the selected sites which are from two different regions – northeast and midwest. In Section 2 I present both daily and monthly time series plots which would roughly show the temporal variability and trends if any. Also, in Section 2, I present the comparison of variability between two different regions and within region. In Section 3 I introduce the theoretically mathematical relationship between pH and rainfall and present the algorithm to generate the parameters of nonlinear models and the results. I also present the residual plots to indicate the aptness of the given model and the stability plot of the parameters of the model. In the last section I present the summary and what further analysis may be needed.

2. Trends and Seasonal Patterns.

This section presents the daily time series plots and monthly time series plots to see if there are any annual trends or seasonal patterns.

For the daily plots, the X axes represent the days whose range is from 1 to 2000. The first day is Julian day 1, 1979, the second day is Julian day 2, 1979, etc. up to the 1830th day, which is Julian day 366, 1983. Even for 2 years data, I made the range from 1 to 2000
because the same scale would make the differences or patterns easier to see. The Y axes may be rainfall or the natural log of sulfate concentration or ph. Because I have used the same scale, some extremely low or high data points have been truncated. Daily plots present almost every data point we have.

The total rainfall figures are almost all between 0 and 6 centimeters. A few of them are greater than 6 centimeters, so these were truncated. The rainfall time series plots (Figure 2) did not indicate any trend, and it was hard to tell if there was any pattern.

In order to clearly see the distribution of the sulfate concentration, a natural logarithm transformation was made to sulfate concentration. In Figures 3 and 4 there are 6 log sulfate concentration plots. They all showed very clear seasonal patterns. The concentration is higher during the warm season and, conversely, lower during the cold season. Figures 3 and 4 each include three different sites from the midwest and northeast, respectively. As we can see, the day-to-day variability is discernibly greater in the northeast than in the mideast. The range is from 1 to 100 for the northeast, but 5 to 100 for the mideast. Geographically, the annual sulfate concentrations are higher in the midwest than in the northeast.

Because sulfate is one of the strong acids, I would expect the same kind of pattern to occur in the ph time series plot. As we see in Figure 5, the seasonal pattern in the ph plots is not as clear as in sulfate concentration, but it is still discernible. The high and low ph values occurred during cold and warm seasons, respectively. That means there is more acid (low ph value) during the warm season and less acid (high ph value) during the cold season. This is exactly what we expect according to what we found from the sulfate plots.

It is hard to see any trend in ph of the precipitation based on the daily time series plots. But since possible downward trends in the ph have been of great concern, I then looked at the weighted monthly average time series plots.

The ph value is the negative logarithm of the hydrogen ion concentration expressed in
moles per liter. So the weighted monthly average of pH is given as follows.

$$\text{average pH} = -\log \left[ \sum_{i=1}^{30} \frac{10^{-\text{pH}} \times \text{rainfall}}{\text{or during one month}} \right] / \sum_{i=1}^{30} \text{rainfall}$$

where the numerator of the fraction is the total hydrogen ion concentration and the denominator is the total rainfall in one month. The weighted monthly average of sulfate concentration is given by

$$\text{average sulfate} = \sum_{\text{during one month}} \frac{\text{sulfate concentration} \times \text{rainfall}}{\sum_{\text{during one month}} \text{rainfall}}.$$

The monthly lots (Figures 6, 7) did show exactly the same pattern as what we have seen in the daily plots for both sulfate concentration and pH. Also, the plots showed no observable trend over the 5 year period at 7 sites. Moreover, the plots indicated the highest pH and sulfate concentrations were during August and September and the lowest were during January, February, and March.

3. Estimate of the Parameter of the Mathematical Model of Rainfall and pH.

The data plots (pH against rainfall) are given in Figure 8. Now we would like to see the mathematical relationship between pH and rainfall.

For a simple scavenging model we assume that each raindrop takes a certain proportion of available acid in the air and the second raindrop will carry the same proportion of the acid remaining after the first raindrop. Based on this idea, the relationship between pH and rainfall is given below as in Eynon and Switzer (1983),

$$\text{pH} = \log \left( \frac{\text{rainfall} \times c}{1 - \exp(\text{rainfall} \times c)} \right) + b \quad (1)$$

where $b$ and $c$ are the parameters of the model.

The graph of the above function is given in Figure 9 for different values of $c$. $b$ is the intercept of the curve. As we can see from Figure 9, the greater the $c$ value is, the greater the
change in ph per unit change in rainfall.

Now the problem is that the given mathematical model is nonlinear, so how do we estimate the parameters of the nonlinear model? The following paragraph states the algorithm for generating the estimate of the nonlinear model parameters. The iteration algorithm is an application of Newton's method.

The algorithm is as follows.

(i) Initialized parameter $c$, say $c = c_0$. Then the model becomes linear with parameter $b$.

(ii) Apply the least squares method to estimate $b$, say $b = b_0$.

(iii) Expand the function (1) to the first order of Taylor series at $(b_0, c_0)$. Simplifying the equation we have a linear function with parameters $b$ and $c$.

(iv) Apply the least squares method to the linear model (given in step 3) to estimate $b$ and $c$, say $b = b_0$, $c = c_0$, then go to step 3 until the estimates are good enough.

The linear model was derived as follows:

Let $z$ stand for rainfall. Then writing (1) as a function of the parameters $(b), (c)$ we have

$$F(b, c) = \log \left( \frac{cz}{1 - \exp(-cz)} \right) + b.$$ 

The first order Taylor expansion is given as follows

$$F(b, c) \approx F(b_0, c_0) + \frac{\partial}{\partial b} F(b, c) \bigg|_{(b_0, c_0)} \cdot (b - b_0)$$

$$+ \frac{\partial}{\partial c} F(b, c) \bigg|_{(b_0, c_0)} \cdot (c - c_0).$$

In order to get the expansion, we need to calculate the first order derivatives with respect to $b$ and $c$. Since

$$\frac{\partial}{\partial b} F(b, c) = 1$$

and

$$\frac{\partial}{\partial c} F(b, c) = \frac{1}{\ln(10)} \left[ \frac{z \cdot \exp(-zc)}{1 - \exp(-zc)} + \frac{1}{c} \right].$$

We have

$$F(b, c) = Y = \log \left[ \frac{c_0 z}{1 - \exp(-c_0 z)} \right] + b_0 + 1 \cdot (b - b_0)$$

$$+ \frac{1}{\ln(10)} \left[ \frac{z \cdot \exp(-zc_0)}{1 - \exp(-zc_0)} + \frac{1}{c_0} \right] \cdot (c - c_0).$$
Simplifying the above equation, we have

\[ Y = \log \left( \frac{\frac{zx_0}{1 - \exp(-zx_0)}} {\ln(10)} \right) + \frac{1}{\ln(10)} \left( \frac{z \cdot \exp(-zx_0)} {1 - \exp(-zx_0)} + \frac{1}{c_0} \right) \cdot c_0 \]

which is a straight line equation with intercept \( b \) and slope \( c \).

Regressing the dependent variable

\[ Y = \log \left( \frac{\frac{zx_0}{1 - \exp(-zx_0)}} {\ln(10)} \right) + \frac{1}{\ln(10)} \left( \frac{z \cdot \exp(-zx_0)} {1 - \exp(-zx_0)} + \frac{1}{c_0} \right) \cdot c_0 \]

on the independent variable

\[ \frac{1}{\ln(10)} \left( \frac{z \cdot \exp(-zx_0)} {1 - \exp(-zx_0)} \right). \]

We then have the least square estimates of the intercept \( b \) and slope \( c \).

Applying the algorithm above to 14 sets of data, 7 sites and 2 seasons (winter and summer), we have the results listed in Table 1.

Table 1

<table>
<thead>
<tr>
<th>Region</th>
<th>Site #</th>
<th>Site #</th>
<th>Site #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Winter</td>
<td>Summer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Parameter</td>
<td>Parameter</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( b )</td>
<td>( c )</td>
</tr>
<tr>
<td>7</td>
<td>4.56</td>
<td>-0.32</td>
<td>4.29</td>
</tr>
<tr>
<td>Midwest</td>
<td>4</td>
<td>4.23</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>4.10</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21</td>
<td>4.24</td>
</tr>
<tr>
<td>Northeast</td>
<td>20</td>
<td>4.21</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4.07</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>4.10</td>
<td>2.36</td>
</tr>
</tbody>
</table>

From Table 1 we found that the \( b \) values are greater for the winter season than the summer season. That implied that there is less acid during the winter season. Also, most of the \( c \) values are greater for the winter than the summer. This implies that the pH dependence on rainfall
volume is greater rate for the winter than for the summer. Also, most of the $c$ values are greater for the northeast than the midwest.

Now we would like to examine the aptness of the given model. By taking a look at the residual plots (Figure 11), we see that the residual variance is decreasing in rainfall, i.e. the residual plots show the lack of homogeneity. So an appropriate transformation of variables in the model may be needed in order to stabilize the error variance.

Next we would like to examine the stability of the estimate of the parameters. Does this mean that a different set of data taken from the same site during a similar period will give "similar" estimates of $b$ and $c$? To do this, we apply the bootstrap method. Let $N$ be the number of a set of data points. Draw $N$ times from $\{1, 2, \cdots, N\}$ according to a multinomial distribution with equal probability $1/N$ for $1, 2, \cdots, N$. This results in a set of indices $i_1, i_2, \cdots, i_N$. The corresponding bootstrap sample is $t_{i_1}$th, $t_{i_2}$th, $\cdots$, $t_{i_N}$th points of the data set. The results for three different sets of data are given in Figure 12, which indicates the possibility of substantial sampling variability in the parameter estimates.

As we can see, there seems to be a linear relationship between estimates of $b$ and $c$. The greater the $b$ is the smaller the $c$ is. This is easy to explain. Since we know the error variance is decreasing in rainfall, the data generally look like Figure 10. For area 2, the data points should be from region C. But for area 1, the data points could be more from region A than region B or conversely. If more are from region A we will have higher intercept ($b'$), so to reach area C, only a small increasing rate is needed. Conversely, for a lower intercept ($b''$) a greater increasing rate is required.


This section will give the summary of what has been seen and what further analysis may be needed. What we have seen is listed below.

(1) both daily and monthly time series plots for 7 selected sites from 2 different regions indicated no discernible trends for the pH of the precipitation.

(2) There was no clear seasonal pattern for the rainfall, but both daily and monthly plots did show clearly that there is a seasonal pattern for pH and the log of sulfate concentration.
high sulfate concentration – warm season

low sulfate concentration – cold season

high pH (less acid) – cold season

low pH (more acid) – warm season

Since sulfate is one of the strong acids, the seasonal pattern for sulfate concentration and pH are expected to be similar. The plots showed this.

(3) The seasonal patterns of sulfate concentration are exactly the same for the 7 selected sites, except the day-to-day variability is greater in the northeast than in the midwest.

(4) For a rainfall scavenging model residual plots showed lack of homogeneity so appropriate transformation of the variables in the model may be needed in order to stabilize the error variance.

(5) The scavenging model parameter estimates vary by season and location and are not always well estimated from the available data.

What further analysis may be needed? Of course, a lot of different data analysis will be very interesting as a continuation of this report. The following analysis will help with a more detailed conclusion.

(1) Comparison of year-to-year and site-to-site variation.

(2) More geographical comparisons.

(3) Appropriate transformation to the variables in the model to stabilize the error variance.
Figure 1  Utility Acid Precipitation Study Sites
Figure 4

SITE NO. 20

SITE NO. 1

SITE NO. 21
Figure 6
Figure 10  Interpretation of the Relationship
Between Parameter Estimates $b$ and $c$