SERIAL CORRELATION OF MONTHLY SNOWFALL IN NEW ENGLAND

by
William E. Reifsnyder and George M. Furnival[1/]
Yale University

Introduction

At the 1960 meeting of the Eastern Snow Conference, Jerome Namias discussed factors affecting the monthly and seasonal variations in snowfall over the northeastern United States. In this paper (Namias, 1960), he attempted to analyze the flow patterns that were associated with winters characterized by varying amounts of snowfall. It was his expressed hope that the study would provide some of the groundwork basic to a rational attempt to predict the character of winter.

As the area of interest, Namias took the ten northeastern states from Ohio to Maine. Within this area, he selected 32 stations to provide a basis for determining the average monthly snowfall. He then computed average snowfall for the months of December, January, and February, as well as the cold season total for the winters 1929-30 through 1958-59. Upper level flow patterns were determined from mean winter charts of the 700-millibar height contours.

After ranking months and winters in terms of snowfall amounts, Namias proceeded to examine the upper level flow patterns associated with those months and seasons with higher- or lower-than-normal snowfall. We would like to quote in particular one paragraph of his discussion relating to these flow patterns:

> "The impression one obtains through study of many charts such as these and related material, is that an anomalous upper level wind pattern becomes established for the winter and persistently recurs in similar form with only brief transitory deviations. These anomalous wind patterns to a large extent set the stage for certain types of storm development in particular areas, and moreover, because of their ability to deploy warm or cold (and wet or dry) air masses into the storm areas, determine the growth rate of the storm once formed. The course of the storms is also frequently determined by these quasi-stable wind patterns, which in effect steer them. Thus the central problem involved in winter snow characteristics, and indeed in all other types of climatic fluctuations, must involve the prevailing wind patterns of the general circulation of the atmosphere."

Present Study

This analysis of Namias suggested to us the following reasoning and hypothesis. If flow patterns become established for the entire winter or a significant portion of it, and if these patterns are associated with snowfall patterns and amounts, then the snowfall amount in any one month should be correlated with that of the previous month or months. If such a correlation appeared, it would then be possible to develop regression equations that would provide better predictions of monthly snow fall amounts than would be possible using climatic averages. In other words, we wished to test Namias' conclusions and, if possible, put them in form that could be readily used to obtain predictions of snowfall that were better than climatology.

[1/]Associate Professor of Forest Meteorology, and Associate Professor of Forest Mensuration, respectively, Yale School of Forestry, New Haven, Conn.

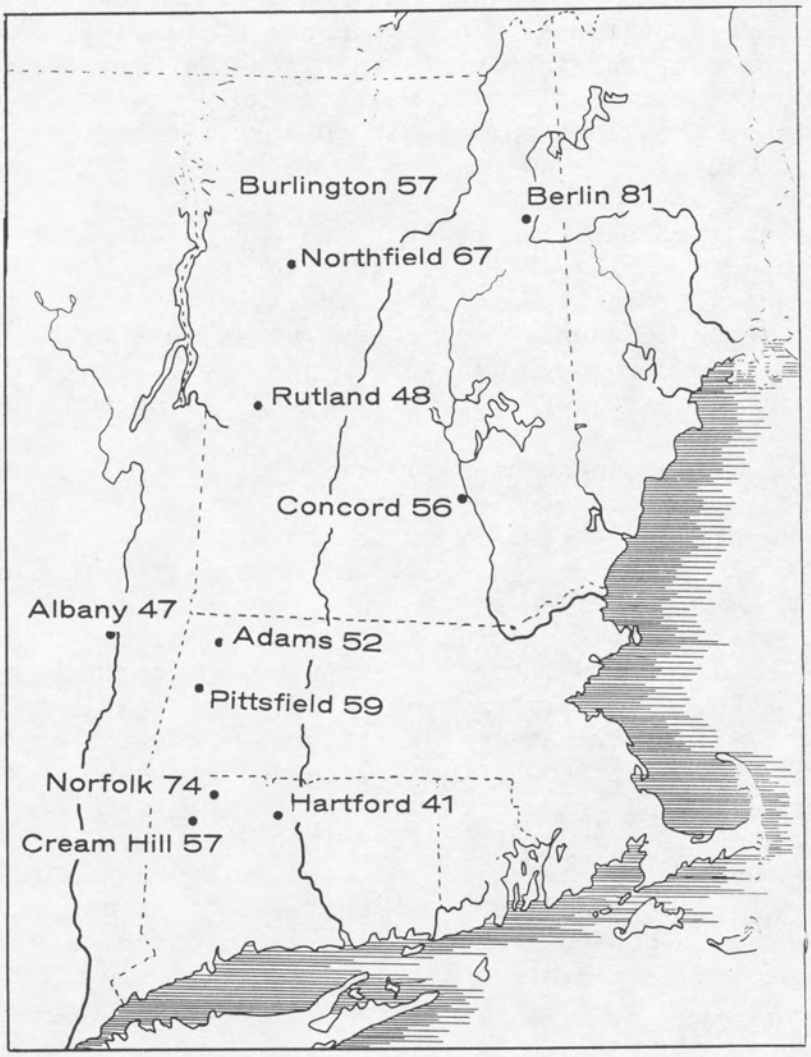


Fig. 1. Stations used in analysis. Numbers are 30-year mean snowfall in inches.

There are a number of difficulties in using statistics to prove hypotheses about weather phenomena. Because of their importance for this study and for others of a similar nature, we should like to discuss them in some detail.

First of all, it is absolutely necessary that the hypothesis and the statistical procedures used to test the hypothesis be established before the data used in the analysis is inspected or analyzed in any way. It is not valid to study the data for possible relationships, then use the same data in order to test whether or not the relationships are valid. A brief and simple example will demonstrate the importance of this.

Suppose that we have a population in which two variables or attributes are distributed randomly. Now let us draw sets of samples from this population, each sample consisting of a pair of observations. Now if we accept the 95% level as showing significance, we will find that one out of 20 sets will show a "significant" correlation. This correlation is of course spurious and occurs only by chance. If one draws a hundred sets of samples and then tests the five or so sets that appear to show a correlation, he may be misled into thinking he had discovered a real correlation, whereas he is actually dealing with a random population.

In our study the statistical procedures were in fact established without reference to the data used in the analysis. We decided to test the null hypothesis that there was no correlation between snowfall amounts in a limited area between the four winter months of December, January, February, and March. In order to determine the snowfall for the area, we decided to use a number of stations for which records were available for a suitable number of years.

This brings us to the second major consideration in the statistical analysis. The question to be answered can be stated as follows: should we consider our sample as being a collection of snowfall amounts occurring during years that are essentially independent of each other and therefore can be considered as random, or could the individual stations be considered the sampling unit? Unfortunately, since snowfall amounts at nearby stations are highly correlated, the addition of one station to those previously selected does not add greatly to the total amount of information on areal snowfall amounts. Because of this correlation, it is not valid to consider the additional station as an additional independent sample.

On the other hand, the addition of one year of data does add greatly to the information because the events of one winter can be considered to be completely independent of the events of the year or years previously. The test of signifcance must consider years as the replicated variable. The information contained in the data for individual stations is not lost, however. Obviously, what we are interested in here is obtaining an estimate of snowfall amount over a restricted area. Because of local variations, the snowfall amount at a single station is sometimes not representative of a very large area. Therefore, additional stations may be added to the analysis in order to obtain a better or more stable estimate of the areal snowfall.

Data and analysis

The area we chose as being appropriate for a correlation analysis of monthly snowfall amounts comprises the central and southern New England area (Fig. 1). We selected eleven stations with long and consistent snowfall records to characterize the snowfall of the area (Table 1). The time period started with the winter of 1930-31 and ended with the winter of 1959-60. The length of record is thus thirty winters, overlapping the period of Namias' study except that his period started and ended one year earlier. We chose the four months of December, January, February, and March as constituting the winter snowfall

period.

We used a smaller area than that of Namias for several reasons. If correlations exist on areas the size of a watershed (the Connecticut River, for example), then the prediction equations would be useful for estimating total snowpack available for spring melt. It was thought further that the smaller area would be somewhat more homogeneous than the entire Northeast with regard to the origin and passage of winter storms, and that such a selection might reduce the variance of sample snowfall amounts. Also, we did not wish to use the same data that Namias used, in order to avoid the pitfall mentioned earlier -- that of developing a hypothesis from data, then using the same data to "test" the hypothesis. It would have been better to use a different series of years, or a completely different area. This was not possible, so we did the next best thing: we used a smaller area, and different stations (except for several duplications), and one new year of data.

Table 1

New England States Used in Snowfall Analysis

Station	Elevation (feet)	30-yr. mean snowfall Dec. - March (inches-depth)
Burlington, Vt.	321	56.97
Northfield, Vt.	840	67.20
Rutland, Vt.	600	48.11
Berlin, N. H.	1110	80.55
Concord, N. H.	339	55.66
Albany, N. Y.	277	47.08
Adams, Mass.	800	52.36
Pittsfield, Mass.	1000	58.81
Norfolk, Conn.	1380	73.67
Cream Hill, Conn.	1300	56.98
Hartford, Conn.	63	40.67
Average		58.01

The analysis was a straight-forward test of the independence of a set of variates. A matrix of correlation coefficients was obtained -- each month being paired with each other month in the study. The hypothesis tested is that the matrix of correlation coefficients is not different from the unit matrix. The procedure used is that described by Anderson (1960) in which a transformation of the determinant is distributed approximately as Chi-square:

TEST OF SIGNIFICANCE OF CORRELATION MATRIX

$$\chi^2_f = - m \log_e R$$

where $f = \frac{p\,(p - 1)}{2}$

$m = N - \frac{2p + 11}{6}$

N = number of independent sets of observations

p = number of variables per observation

R = determinant of correlation matrix

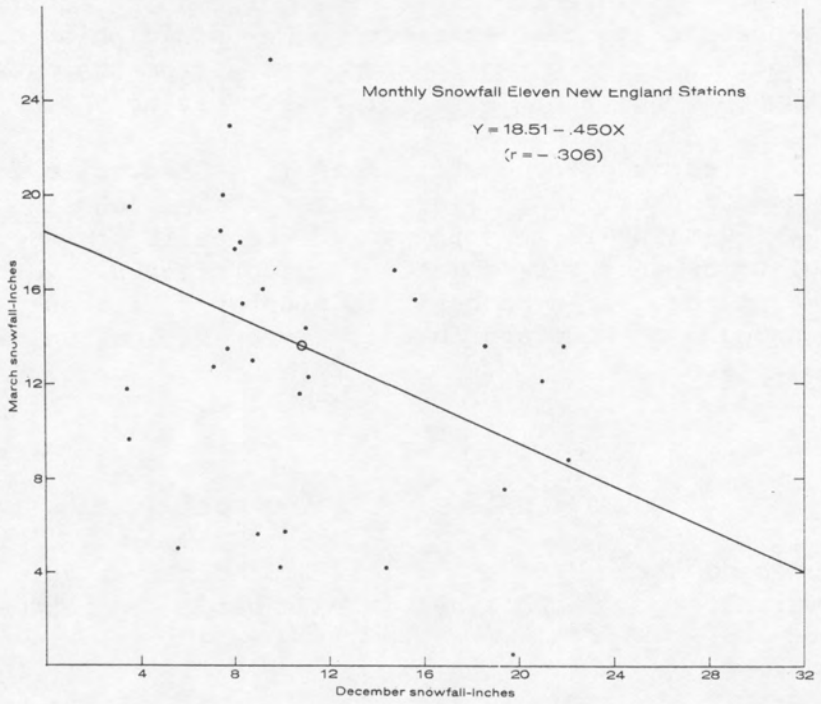


Fig. 2. March snowfall as a function of December snowfall.

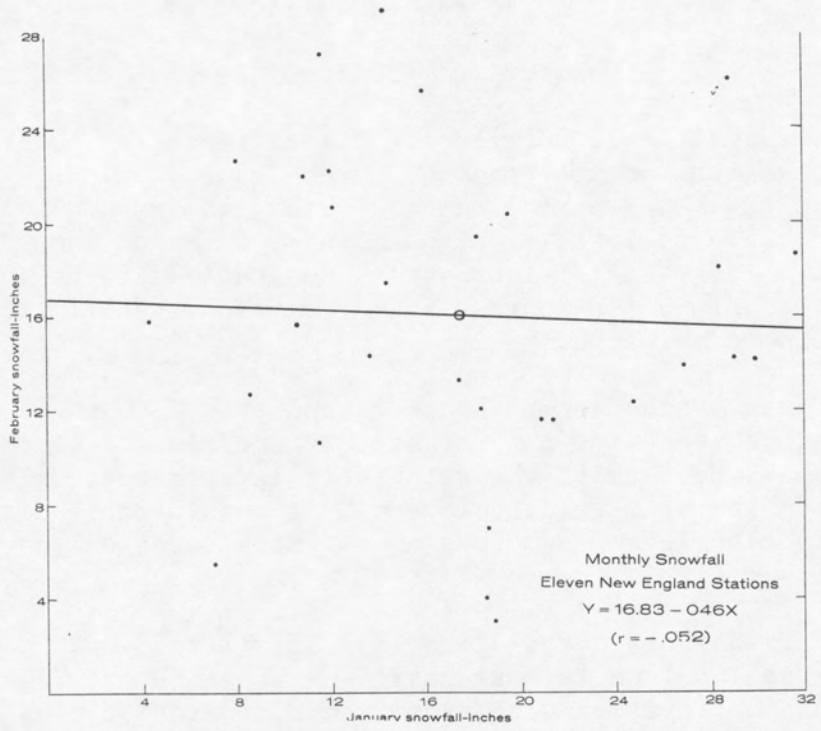


Fig. 3. February snowfall as a function of January snowfall.

In the absence of any independent hypothesis about the relative merits of any particular pair or pairs of months, it was necessary to restrict the significance test to the entire matrix. Obviously, in a sample of randomly distributed pairs, some would turn up "significant" by chance alone. We felt that it was not valid therefore to inspect the correlation matrix first, then test the ones that appeared to be significant. We should point out, though, that it is perfectly permissible to develop hypotheses from the data; it is just that the new hypotheses require new data for their testing.

The matrix of correlation coefficients is presented in Table 2. According to the test used, described earlier, there is about one chance in five that a larger chi-square would be obtained if the pairs of observations were drawn from populations that were completely uncorrelated. There is thus not much evidence that the null hypothesis is disproved, although the sample chi-square is large enough to encourage further investigation, on different populations of snowfall data.

Table 2

Matrix of correlation coefficients

Independent variable	Dependent variable: December	January	February	March
December	1	+ .152	+ .168	- .306
January		1	- .052	- .409
February			1	+ .047
March				1

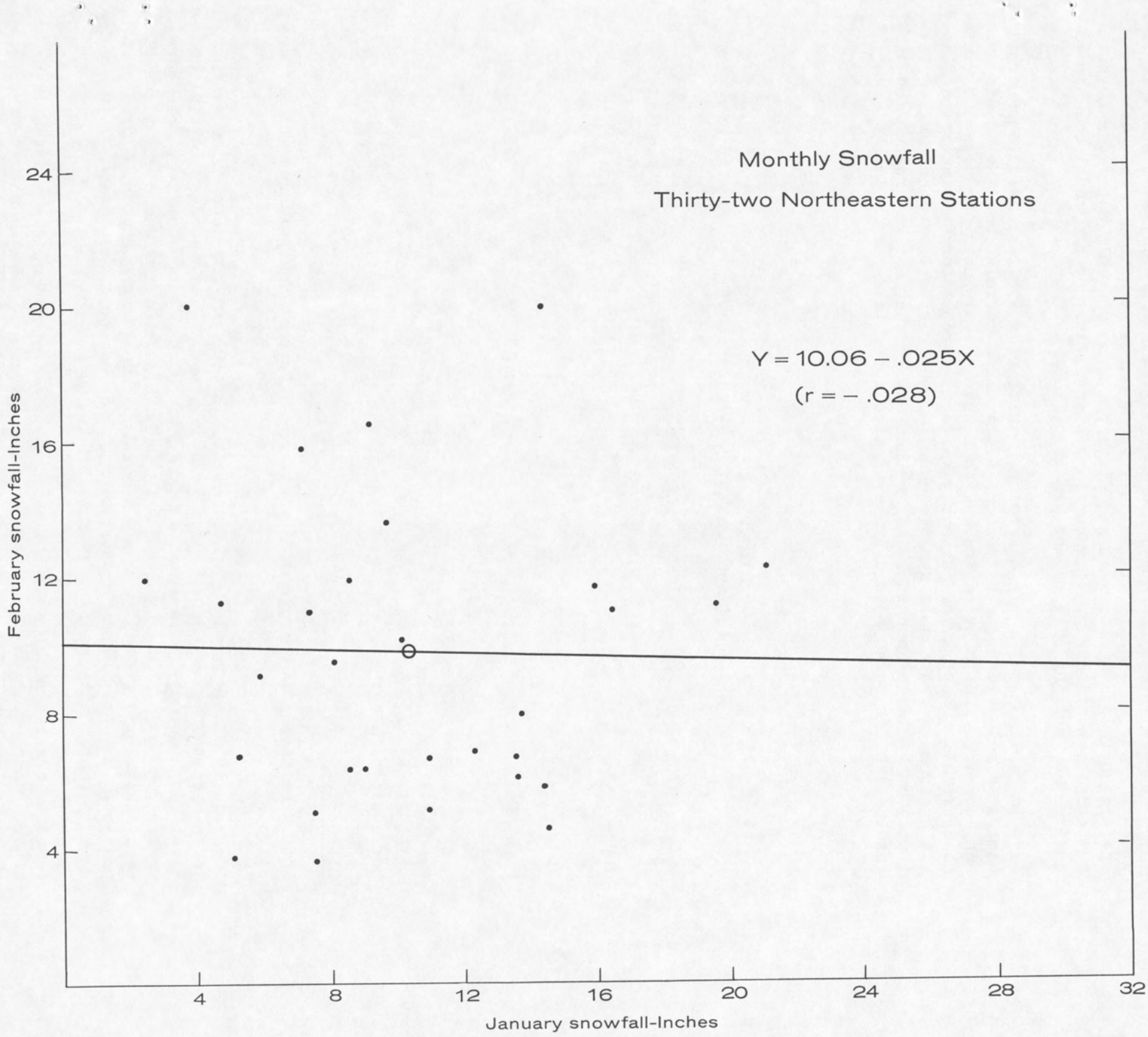
Computed sample chi-square: 8.682

Chi-square, .20 prob. level: 8.558

Some of the individual correlation coefficients are moderately large, for example, the March-December and the March-January values. Surprisingly, they are both negative, implying that, if real, low snowfall amounts in December and January are followed by high amounts, in March, and conversely. Does this mean that a large-scale flow pattern does not last the entire winter, but will persist for only a few months, then reversing itself? The data are suggestive, but far from conclusive.

A plot of the March-December data is shown in Figure 2. This one has the largest slope, that is, the largest "b" coefficient of the six possible simple regressions. Because of the relatively large slope, an estimate would usually differ from the climatological mean by a considerable amount. However, the scatter is also large, and thus an estimate based on the relationship would not be very reliable.

The February-January correlation is nearly zero, and the slope of the calculated regression line is very nearly parallel to the X-axis (Fig. 3). Thus, even if the correlation were significant, there would be very little improvement over the climatological or 30-year mean if the regression equation were used. For the entire range of January snowfall, from four to thirty-two inches, the "estimate" of the February snowfall varies less than two inches. Thus we are forced to conclude that, with the data used, the correlations are not demonstrably significant; and even if they were, the predictive value of the regression equations would be slight in four of the six pairs of months.


Monthly Snowfall
Thirty-two Northeastern Stations
Y = 10.06 – .025X
(r = – .028)
February snowfall-Inches
January snowfall-Inches
4
8
12
16
20
24
28
32

We next looked at a subsample of our own data -- the five northernmost stations, those in Vermont and New Hampshire. The results of this analysis are presented in Table 3, along with the results for the eleven-station analysis. As one would expect, there is little difference between the 5-station and the 11-station analysis. It appears, therefore, that the five northern stations do not behave much differently from the six more southerly stations. It is interesting to note that there is only a one-out-of-ten chance that the five-station correlation matrix is not different from the unit matrix, and it is tempting to ascribe significance to this analysis. We are cautious, however, and are not willing to take the plunge.

Also in Table 3 and Figure 4 are the results of a similar analysis performed on Namias' original data. We found no significant correlation. We should point out, however, and in strong terms, that our analysis does not <u>disprove</u> Namias' conclusions. They may be perfectly valid; a more elaborate statistical analysis based on a reasonable dynamic model, might very well demonstrate the validity of Namias' reasoning. All we can say is that <u>our</u> interpretation of his conclusions is not supported by <u>our</u> analysis of his data.

Table 3

Correlates				
Y	X	Northeast (32 stations)	New England (11 stations)	N. New England (5 stations)
Jan.	Dec.	+.242	+.152	-.055
Feb.	Dec.	+.230	+.168	+.003
Feb.	Jan.	-.028	-.052	+.152
Mar.	Dec.		-.306	-.257
Mar.	Jan.		-.409	-.465
Mar.	Feb.		+.047	+.048
Probability of larger value of chi-square, correlation matrix		.35	.19	.10

<u>Discussion</u>

What does this study prove or disprove? Is further analysis feasible or desirable?

The results are equivocal -- they are suggestive but not conclusive -- they lie in an uncertain area between proof and disproof. Certain ideas for further analysis and testing arise from the data, but how to test them? We have very nearly run out of data. We can stretch back a few years and pick up earlier data for several of the stations. But this runs into the problem presented by possible long-term climatic changes that may confound the data. Or, we can wait another thirty years to accumulate a new set.

There is a substantial question as to whether refinements in hypothesis or statistical procedure can be tested on these data. In a sense we may have spoiled the data for future analysis based on the present work. This is often a major obstacle in the use of the climatic record in statistical analysis.

Another difficulty in analyses of the sort we have engaged in stems from the tremendous inherent variability of meteorological data. Because of this variability, modelling of a relatively unsophisticated nature, such as we have done here, is likely to have so much variation that extremely large samples are required to provide an adequate test of significance. Even if significance can be demonstrated, the large amount of unexplained variation leads to coarse and inefficient regression estimates. In climatological work, refinement usually means highly sophisticated models and statistical techniques, and data stretching for many years.

Acknowledgements

We should like to acknowledge the assistance of Joseph Brumbach, State Climatologist for Connecticut, in obtaining the basic data for the study; and the assistance of Richard Couser, Yale University, in performing the statistical computations.

Literature cited

Anderson, Theodore W. 1960. An introduction to multivariate statistical analysis. New York, John Wiley. 374 pp.

Namias, Jerome. 1960. Factors leading to variations in monthly and seasonal snowfalls over Eastern United States. Proceedings, Seventeenth Annual Meeting, Eastern Snow Conference, Feb. 4 and 5, 1960. P. 167-189. Also published in Weatherwise, 13(6): 238-247.