Report SDSMT/IAS/R04-01 February 2004

EXPLORATORY ANALYSIS OF CLIMATIC RAINGAGE DATA FOR EVIDENCE OF EFFECTS OF THE NORTH DAKOTA CLOUD MODIFICATION PROJECT ON RAINFALL IN THE TARGET AREA

By: Paul L. Smith
Paul W. Mielke, Jr.
Fred J. Kopp

Prepared for:

North Dakota Atmospheric Resource Board
900 East Boulevard Ave., Dept. 410
Bismarck, ND 58505-0850

Contract No. ARB-IAS-02-1

Institute of Atmospheric Sciences
South Dakota School of Mines and Technology
501 East Saint Joseph Street
Rapid City, SD 57701-3995

1. Introduction

The primary objective of the North Dakota Cloud Modification Project (NDCMP), which has operated in western North Dakota since 1976, is to reduce losses due to hail damage. Stimulation of enhanced rainfall is a secondary objective, and in any case knowledge of hail suppression effects on total precipitation would be important to the project's operation. This report summarizes an exploratory analysis of rainfall data for the NDCMP target areas and an upwind control region in eastern Montana, intended to elucidate any such effects.

The planned approach was to use techniques similar to those employed in Smith *et al.* (1997) to explore the effects of the NDCMP seeding on crop-hail insurance losses. In brief, the approach uses data from the target and from upwind control areas, for both the seeding period of interest and a prior historical period, in a statistical analysis based on the multi-response permutation procedures (MRPP; Mielke and Berry 2001). Techniques like those described in Smith *et al.* can provide estimates of the probability that any observed differences could be due to random chance, along with point and confidence-interval estimates for any difference that may be found statistically significant.

Two differences from the earlier hail insurance data analysis procedure were initially considered useful. One is improvements in the statistical procedure to use multivariate multiple least-absolute-deviation (LAD) regression techniques (Mielke and Berry 2002). The other is to subdivide the control area, consisting of the same 12 counties in eastern Montana used in the hail insurance analysis, into a north control and south control. That would permit two separate target versus control analyses, yielding some indication of the stability of the results, along with a north versus south analysis which would give some sense of the possibility of finding a "seeding effect" in these areas where no seeding in fact occurred.

2. Climatic raingage data

The North Dakota Atmospheric Resource Board (NDARB) operates a statewide network of some 900 reporting stations (rain and hail) during its summer – usually June through August – operating season. This provides a mean gage density of about one gage per 200 km^2, and data which would have obvious value because of the close spacing. Unfortunately, neither upwind control-area data from Montana nor historical data prior to the inception of the NDCMP are available from this network. Consequently the NDARB data do not lend themselves to the type of analysis carried out here. Scientists at the University of North Dakota are carrying out a separate analysis of the NDARB data, and it will be reported elsewhere.

Climate stations operated as part of the NOAA climatic network measure rainfall along with other quantities of interest. A few of these stations in the NDCMP target area have records dating back at least as far as 1894, but the earliest Montana data are from 1911. The data from this network were chosen as the basis for this analysis; we focused

on the monthly rainfall totals, as that could provide indications of possible seasonal variations.

A total of 11 stations in the NDCMP target areas and 25 stations in (10 of the) 12 easternmost counties of Montana have remained in the same location and currently report rainfall data as part of the climatic network. An additional gage at Marmarth, in western Slope County, and two gages in Hettinger County, North Dakota, could not be used because those areas have dropped out of the NDCMP in recent years. Unfortunately, the period of record varies greatly; while records for a few of the stations go back to 1894, others only began reporting in 1950 or even later. Moreover, one station in the target area is lacking data for 15 years of the NDCMP era, while one in the (north) control area is lacking 20 years of data from the 1950s to 1970. Table 1 illustrates the decrease in the number of available stations with reasonably complete records as the period of interest is extended further back in time.

Table 1: Availability of Climatic Rain Gage Data

Start Year	Number of Stations with Usable Data	
	Target	Control (N/S)
1950	10	21 (10/11)
1949	10	16 (8/8)
1948	10	11 (6/5)
1946	9	11 (6/5)
1942	8	10 (5/5)
1938	7	10 (5/5)
1936	6	8 (4/4)

To begin the analysis with an earlier start date would provide a longer period of record with no seeding in the area (there has been some cloud seeding activity in western North Dakota most years since 1951) but a smaller number of gages with the requisite data continuity. After review of the information in Table 1, we decided to begin the analysis with data from 1950. Trying to extend back to include even two more years would lose almost half of the control-area stations. This choice of the analysis period does involve a risk of diluting possible indications of any seeding effect, in view of the seeding activities in the target region since 1951. However, there appeared to be no viable alternative as the number of available gages even from 1950 is already small for obtaining good estimates of the rainfall in the area.

The definition of the "target area" for this analysis also warrants some discussion. About District 2 there is little question; all three counties (McKenzie, Mountrail, and Ward) have participated continuously in the NDCMP since its inception. However, there has been considerable variation in the makeup of District 1. Hettinger County was part of the district only from 1976 through 1988, so was not included in this analysis because that is barely half the NDCMP time period. Bowman County participated in all years

except 1990, and so was included in the analysis – treating 1990 as among the "seeded" years, even though no seeding actually occurred. Slope County was out of the program in 1990 and again in 1999, and since 2000 only the eastern part of the county has participated. Consequently only the eastern part, containing two climatic gages, was included in the analysis, on the same basis as was Bowman County.

2.1 Adjustment for missing data

Beyond the difficulties already noted, the climatic gage records include three types of entries in addition to the actual gage readings. These are notations concerning "missing", "incomplete", and "estimated" data. The "incomplete" values represent cases where data from a few days during the month, evidently a maximum of nine, are lacking; those values were incorporated as is into the analysis, there being no reason to anticipate any systematic bias resulting from their occurrence. "Incomplete" values affect some data from about one-third of the stations, but only one station had more than four "incomplete" months (out of the total of 159 months included in the analysis). In that instance, all nine "incomplete" months were August data, a period of generally light rainfall in the region and possibly when the observers tended to take vacation. Overall, there were 13 "incomplete" gage-months in the target area, out of a possible 1,590 gage-months (10 gages over 3 months for 53 years) – that is, less than 1%. For the control area the comparable numbers are 15 incomplete gage-months out of a possible total of 3,339 – less than 0.5%.

The process for arriving at the "estimated" values is not known to us, but as it is evidently satisfactory to the National Weather Service we also used those values as is in the analysis. All the "estimated" values were for control-area gages, with 22 gage-months affected (less than 0.7%). Once again a single station was responsible for more than half of those cases.

"Missing" monthly values appeared in the records for most of the stations, with up to six monthly values missing for any one station. For the target area there were 20 "missing" gage-months out of 1,590 possible, or 1.26%, while there were 31 "missing" out of 3,339, or 0.93%, for the control area. Whereas the "incomplete" and "estimated" values were used directly in the analysis, some means was needed to replace the "missing" entries with values that represented plausible estimates of missing data.

For that purpose, area by area, station by station, and month by month multiple LAD regression relationships were used. That is, for missing June 19xx data for Station A in the target area, June data for all target stations and for all other years with June data were used to develop a regression relationship to predict the June values for Station A. Then the existing data values for the other stations for June 19xx were used as predictors to obtain an estimate for the missing value for Station A. In two, out of the 51 total, instances the predictand value turned out to be negative, so the entry in the database was set to 0.

To illustrate the impact of this process, consider the case of the July data for "Watford 2" station in McKenzie County in NDCMP District 2. The records showed data missing for four Julys out of the 53 years of interest. The range of observed values from the other years was 0.83 - 7.68 inches*, with a median of 2.96 inches (mean 3.32 inches). The regression relationship for those 49 years yielded predictand values ranging from 0.99 to 5.78 inches, in the ratio of from 0.29 to 3.40 times the observed value for the year in question, and involving more than twice as many underestimates as overestimates. For the four missing years the regression estimates were 1.84, 1.92, 3.75, and 6.21 inches (average 3.43 inches). These values indicate regression estimates generally in line with the observed values, and suggest that application of this procedure to about 1% of the total database would not strongly influence the statistical results.

2.2 Test variables

The Appendix lists those stations finally included in the analysis. With the small numbers of gages available to represent the rainfall in such large areas (the NDCMP target area exceeds 20,000 square kilometers and the Montana control area is more than twice as large), meaningful estimates of the areal rainfall volumes are impractical. Consequently, the simple sums of the gage data for each respective area and time period were used as the test variables for this analysis.

3. Characteristics of the data

This section summarizes general characteristics of the primary (season total) data. Table 2 presents some statistical features (medians, means, standard deviations, and correlations) of the various data subsets.

The measures of central tendency, considered on a per-gage basis, reflect the general gradient of precipitation in the region, increasing from west (control area) to east (target area), as well as the generally drier conditions in southeast Montana (south control). No significant temporal trends appear in the data, though the indication of a downward trend for the south control approaches a significant level (a correlation of 0.229 would be significant at the 95% level). This suggestion of a downward trend for the south control, coupled with the weak indication of an upward trend for the north control, may be responsible for an interesting suggestion of a "seeding effect" when the August data from these two non-seeded areas are compared (Section 5).

* Inch units are used in this report because the NOAA rainfall data are recorded in those units; 1 inch = 25.4 mm.

Table 2: Statistical characteristics of the Primary Climatic Gage Data (Values represent June-August season totals in inches.)

	Target (10 gages)	Control (21 gages)	North Control (10 gages)	South Control (11 gages)
Median	73.79	125.06	66.09	62.39
Mean (per gage)	74.405 (7.44)	126.829 (6.04)	65.271 (6.53)	61.558 (5.60)
Std. Deviation	19.43	35.83	18.57	19.80
Coefficient of Variation	0.261	0.283	0.284	0.322
Correlation with Time	0.005	-0.087	0.069	-0.221
Correlation with Target Area	----	0.771	0.761	0.683
Correlation with North Control	---	---	---	0.744

The correlations between areas are more or less typical of this kind of data. They suggest that ordinary linear regression would account for about half of the variance in the predictand data. It may, however, be worth noting that a randomized-crossover rain enhancement experiment in Italy (List *et al.* 1999) with similar correlation between areas did not identify a significant seeding effect.

Interestingly, the overall frequency distributions of the data (Figure 1) suggest that they can be represented fairly well by normal distributions – except for the "outlier" very wet year of 1993. This could be useful in simulations of seeding experiments or other similar studies.

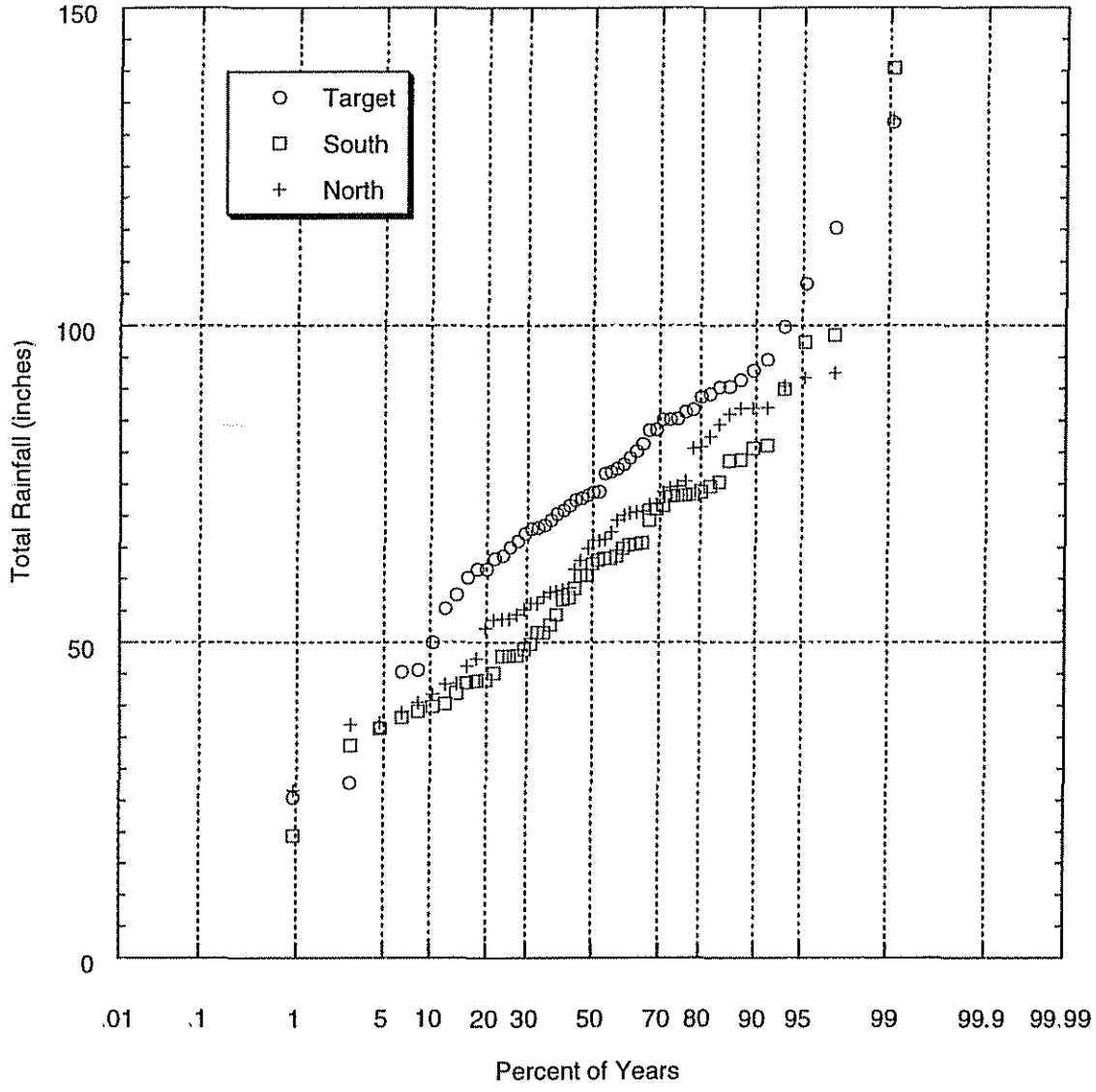

Figure 1: Frequency distributions of season-total rainfall in NDCMP target and Montana control areas, plotted on a normal-probability scale (where a straight line suggests a normal distribution).

4. Primary statistical analysis

The primary analysis procedure begins with a multidimensional scatter plot of the seasonal gage rainfall data. In essence, a point in 31-dimensional space represents the data for a given year; the 31 dimensions represent the respective ten target-area gages and 21 control-area gages. A multidimensional LAD regression line is determined for the points for the entire 53-year data set. We can only represent two dimensions on paper; Figure 2 shows the scatter plot comparing total target-area and control-area rainfall values for the 53 years, along with the corresponding two-dimensional LAD regression line. Here the regression line was forced through the origin, to allow calculation of the point estimates and confidence intervals discussed in Sec. 5.

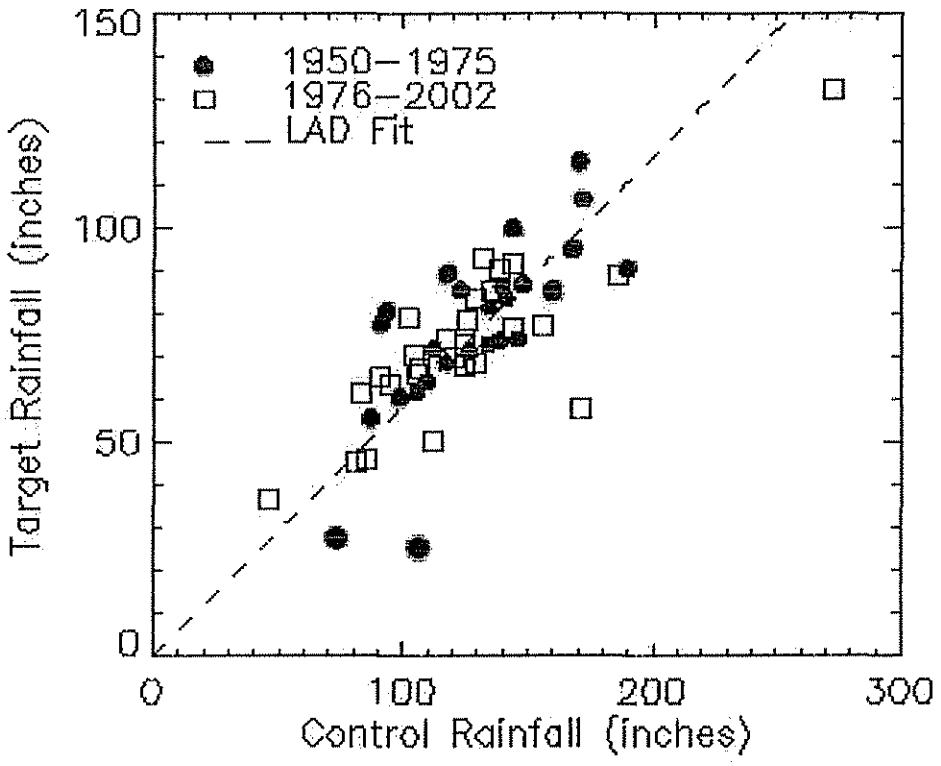

Figure 2. Scatter plot of annual target-total versus control-total rainfall, with corresponding LAD regression line.

Next residual Euclidean distances from the multidimensional line are computed for each of the 53 annual points, and classified into two groups – corresponding respectively to the "historical" years 1950-1975 and the NDCMP years 1976-2002. An MRPP test is then applied to those two groups of residuals to determine whether any difference between the groups is greater than would be likely to occur if the groups of years had been established by random assignment from the set of 53. If the P value (the probability of a test statistic as small as, or smaller than, that actually occurring) under random permutations of the assignments is less than 0.05, or perhaps 0.10, the difference between the two groups of years can be inferred to result from the seeding operations.

The P value for this primary analysis turns out to be 0.322 – i.e. the probability of a test statistic as small as, or smaller than, that actually observed is 32.2%. This result cannot be considered significant by any of the usual measures, so the major result of this analysis is that no significant indication of any effect of the NDCMP seeding on the rainfall in the target area can be identified.

5. Additional exploratory analyses

The accumulated database permits a variety of additional exploratory analyses. At the outset it should be understood that (1) the absence of any significant indication of a primary effect on the seasonal rainfall makes it unlikely that any significant effect will be

found in examining subdivisions, e.g. monthly values, of the data; and (2) the problem of multiplicity in the analyses may lead to some cases having apparently small P values that cannot be accorded the same level of importance as the same P value would have in the primary analysis.

Table 3 summarizes the key results of these exploratory analyses. In brief, none of the MRPP P-values for the various seasonal comparisons is small enough to be considered significant. Among the P values for the month-by-month comparisons, ones for June involving the target area approach the level of 0.10, but in view of the aforementioned multiplicity concerns this cannot be considered statistically significant. Interestingly, the smallest P value (0.069) occurs for the August data in the north control versus south control comparison – suggesting a possible "seeding effect" in the comparison between these two non-seeded areas. All this, of course, merely reflects some combination of the multiplicity factor and the likelihood that the natural variations in the rainfall in the region overwhelm any effect of the NDCMP seeding upon the rainfall as measured by the climatic gage network.

Using Figure 2 and monthly analogs thereof, it is possible to derive point estimates and confidence intervals for any potential seeding effect. The procedure, summarized in the appendix of Smith et al. (1997), uses separate LAD regression lines for the historical and NDCMP years – though in the present instance the large P values do not justify separate lines. The ratio of the slopes of those lines provides the point estimate. Table 3 includes these point estimates, all of which are quite close to 1.0 – a value that would indicate no effect. Confidence intervals shown in the table are obtained with random permutations of the assignment of years to the two groups, followed by recalculation of the LAD regression lines, and indicate the range of the ratio resulting from such permutations. For the seasonal values, the confidence interval is roughly 1.0 ±0.1, while the monthly intervals bracket 1.0 with a somewhat wider range.

Table 3. Key Results of Exploratory Statistical Analysis

Months:	**Summer** (June-Aug)	**June**	**July**	**Aug**
MRPP P values				
Target vs. Control	0.322	0.162	0.828	0.960
Target vs. N. Control	*0.451*	*0.116*	*0.793*	*0.591*
Target vs. S. Control	*0.706*	*0.103*	*0.695*	*0.385*
N. Control vs S. Control	*0.802*	*0.626*	*0.879*	*0.069*
Point estimate of seeding effect	1.008	0.950	0.997	1.020
90% confidence interval	0.91-1.10	0.82-1.10	0.86-1.17	0.88-1.23

The point estimates and confidence intervals for the additional exploratory analyses, for which the P values appear in italics in Table 3, are quite similar for all the season-total comparisons to the primary-analysis values. For the monthly comparisons the confidence intervals are 3 – 5 times wider, and all straddle 1.0. In two cases (August, Target vs. N Control; June, N Control vs. S Control) the point estimate exceeded 1.1, and in one case (August, N Control vs. S Control) it was less than 0.9. In view of the associated P values, discussed earlier, and the wide confidence intervals, no significance can be attached to those values.

6. Summary and Conclusions

This analysis of the climatic rain gage data from the NDCMP target area and upwind control areas in eastern Montana has yielded no significant evidence of an effect of the NDCMP seeding on the summer-season rainfall in the target area. While there may in fact be no such effect, a small effect might not show up in this analysis. For example, an analysis of wheat yield data (Smith *et al.* 1992) suggested an increase of about 6% in the NDCMP target areas that could be attributed to the seeding activity. With the ±0.1 confidence interval indicated in Table 3, even a 6% effect on the rainfall (and part of the wheat-yield effect would reflect reductions in hail losses) would be difficult to find in the climatic rain gage data. Furthermore, the limited number of climate gages available for use in this analysis, the limited usable period of record, and the fact that seeding was taking place in the target area prior to 1976, all may have diluted any indication of a seeding effect under this analysis procedure.

References

List, R., K.R. Gabriel, B.A. Silverman, Z. Levin and T. Karacostas, 1999: The rain enhancement experiment in Puglia, Italy: Statistical evaluation. *J. Appl. Meteor.*, **38**, 281-289.

Mielke, P.W., Jr. and K.J. Berry, 2001: *Permutation Methods: A Distance Function Approach.* Springer, 352 pp.

Mielke, P.W., Jr. and K.J. Berry, 2002: Multivariate multiple regression analyses: A permutation method for linear models. *Psychological Reports,* **91,** 3-9 (Erratum: **91,** 2).

Smith, P. L., L. R. Johnson, D. L. Priegnitz, and P.W. Mielke, Jr., 1992: A target-control analysis of wheat yield data for the ND Cloud Modification Project region. *J. Wea. Modif.*, **24,** 98-105.

Smith, P. L., L. R. Johnson, D. L. Priegnitz, B. A. Boe, and P. W. Mielke, Jr., 1997: An exploratory analysis of crop-hail insurance data for evidence of cloud-seeding effects in North Dakota. *J. Appl. Meteor.*, **36,** 463-473.

Appendix: List of Stations Used

Target Area	North Control	South Control
Amidon*	Culbertson	Baker
Bowman*	Glendive	Belltower
Foxholm	Lindsay	Biddle
Kenmare	Medicine	Broadus
Minot	Plentywood	Ekalaka
Minotex	Raymond	Mackenzie
Reeder*	Savage	Milescty
Stanley	Sidney	Mizpah
Tagus	Westby	Plevna
Watford2	Wibaux	Terry
		Volborg

*District 1 station