United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-83-003b
August 1985

Air

EPA

Evaluation of Rural Air Quality Simulation Models

Addendum B: Graphical Display of Model Performance Using the Clifty Creek Data Base

EPA-450/4-83-003b

Evaluation of Rural Air Quality Simulation Models

Addendum B: Graphical Display of Model Performance Using the Clifty Creek Data Base

Prepared By

William M. Cox
Gerald K. Moss
Joseph A. Tikvart
U. S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Office of Air and Radiation
Research Triangle Park, North Carolina 27711

and

Ellen Baldridge
Computer Sciences Corporation
4501 Alexander Drive
Durham, North Carolina 27709

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air and Radiation
Office of Air Quality Planning and Standards
Research Triangle Park, N.C. 27711

August 1985

This report has been reviewed by the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

PREFACE

This report summarizes performance statistics for several rural point source models based on standardized graphical presentations that allow for both operational and diagnostic evaluations. The performance of the models is evaluated for data collected near the Clifty Creek Power Plant. The report serves as an addendum to a previous publication* on model performance that used extensive statistical summaries in a tabular format as the basis for an operational evaluation. Other addenda to the Clifty Creek publication are also planned for additional data bases and for presentation of further supplemental information on model performance.

*Londergan, R. J. , D. H. Minott, D. J. Wackter, T. Kincaid and D. Bonitata, 1982. Evaluation of Rural Air Quality Simulation Models. EPA Publication No. EPA-450/4-83-003. U.S. Environmental Protection Agency, Research Triangle Park, N.C. (NTIS No. PB 83-182758).

TABLE OF CONTENTS

FIGURES

SECTION 1

INTRODUCTION

The purpose of this report is to provide additional information about the performance of four rural models previously evaluated using the Clifty Creek data base[1]. The goals are two fold: (1) to summarize the statistical comparisons for rural models in a graphical format based on the performance measures already computed and tabulated in the Clifty Creek report; and (2) to provide a framework for development of a standardized procedure for diagnostic evaluation.

The particular data sets and graphical formats shown in this addendum reflect the experience recently gained in presenting and analyzing model performance information. In some cases, data partitions other than those presented in the Clifty Creek report are selected because they appear to provide insight into differences between models that are not easily perceived from the original statistical tabulations. In this analysis, no attempt has been made to infer why a model performs as it does. Complete diagnostic model evaluation is outside the scope since such an evaluation requires additional information and input from the research community.

The four models evaluated are: (1) CRSTER/MPTER developed by EPA; (2) MPSDM developed by ERT, Inc.; (3) TEM-8A developed by the Texas Air Control Board; and (4) PPSP developed by the Martin Marietta Corporation. These models were selected since they span the range of technology represented by available rural models. The reader should refer to the Clifty Creek report and to Addendum A[2] of that report to obtain a more detailed explanation of the models, options used, data bases and data processing procedures.

Figure 1 depicts the relative location of the Clifty Creek power plant to the six monitoring stations which serve as the basis for this evaluation.

Section 2 provides background discussion for the graphics and statistics chosen for presentation. Section 3 provides a general operational comparison of the performance of the four models in selected graphical formats. Section 4 provides a more in depth graphical summary of the performance of each model with results for individual stations and particular meteorological subsets, including dependence on downwind distance and time of day. Hopefully, the data subsets and graphs presented in Section 4 will provide a basis for a standardized approach to diagnostic evaluations that are useful to both the regulatory and model development communities.

The reader should be aware that data bases such as that assembled for Clifty Creek have inherent limitations that must be carefully considered before arriving at general conclusions about model performance. The limitations relate to the representativeness of wind and stability measurements used to characterize atmospheric processes governing plume transport and dispersion. For this site, wind direction and speed were measured at an elevation well below stack height and stability is based on measurements from the Cincinnati National Weather Service Station. Thus, specific results and conclusions presented in this addendum should be viewed as preliminary, pending further analysis with additional high grade data bases.

SECTION 2

STATISTICS AND GRAPHICAL PRESENTATIONS

The American Meteorological Society (AMS) has recommended an extensive variety of statistics and data subsets for use in presenting the performance of air quality models.[3] The original Clifty Creek evaluation, which was patterned after these recommendations, resulted in an overwhelming array of statistical tabulations that were difficult to review and summarize. While subsequent evaluations have been performed using a smaller quantity of statistical output, they have also produced a rather large and unwieldy array of information.[4,5]

Recently, several attempts have been made to focus more closely on selected statistics and data groupings to capture the essential aspects of model performance that are of greatest concern to air quality managers.[6,7,8] In particular, two statistics have been found to be very useful in summarizing and comparing the performance among models. The first statistic, labeled "bias of the average", is calculated as the difference between the average of the observed concentrations and the predicted concentrations. The second statistic, labeled "bias of the standard deviation," is calculated as the difference between the standard deviation of the observed concentrations and the standard deviation of the predicted concentrations. The first statistic measures how well the models estimate the mean of the observed values while the second statistic measures how well the "scatter" of model estimates matches "scatter" in the observed data. For purposes of simplification, the term "scatter" is referred to in place of "standard deviation" throughout this report.

In practice, these two statistics are normalized by dividing by the average of the observed and predicted values. Thus a fractional bias for the average is obtained as

$$FB = \frac{\overline{OB} - \overline{PR}}{(\overline{OB} + \overline{PR})/2};$$

similarly a fractional bias for the standard deviation is obtained as

$$FS = \frac{S_o - S_p}{(S_o + S_p)/2};$$

where $\overline{OB}$, $\overline{PR}$ represent averages and S_o and S_p represent standard deviations for observed and predicted concentrations respectively. The two statistics (FB and FS) can be calculated using any particular data grouping that has relevance. Since these two statistics are used extensively in the following graphical presentations, it is worthwhile to review their properties and interrelationship. The statistic, fractional bias, is mathematically equivalent to (except for a change in algebraic sign) the fractional error used earlier by Irwin and Smith[9].

In Figure 2, the x-axis represents the fractional bias for averages, while the y-axis represents fractional bias for the standard deviation. In each case, a positive bias indicates model underprediction while a negative bias indicates model overprediction. The closer a model is to the center of the figure (i.e., zero bias) the more closely it duplicates the observations. Unlike ratios of observed to predicted values, the fractional bias is restricted to a small finite range. A fractional bias near +2.0 corresponds to a ratio that approaches infinity (∞), for example as predictions approach zero; a fractional bias near -2.0 corresponds to ratios

that approach zero, i.e., as observed values approach zero or when predictions become very large relative to observed concentrations. Also note that a fractional bias between -0.67 and +0.67 indicates average accuracy within a factor-of-two. This factor-of-two range corresponds to the innermost rectangle centered on the origin shown in Figure 2.

While many of the graphs involve the two fractional bias statistics, several other types of graphs are also presented. The Q-Q plot (Quantile-Quantile) is used in Section 3 to compare the 25 highest predicted values with the corresponding 25 highest rank ordered observed values. The information conveyed in the Q-Q plot (e.g., Figure 3) expands on information provided in the fractional bias plot in two important ways: (1) it directly compares the magnitude of the highest individual observed and predicted values, whereas the fractional bias plots are independent of the magnitude of the values; and (2) it highlights changes in the relationship between the predicted and observed values throughout the range of the 25 highest values.

For completeness, cumulative frequency distributions are also presented in Section 3 (e.g., Figure 4). These plots make use of all of the data available, not just the 25 highest values. They illustrate the extent of the discrepancy between predicted and observed values and their degree of departure from a log-normal distribution. In reviewing both the Q-Q and frequency distribution plots the reader should be aware that more than one data point may be represented by a given symbol.

The Figures presented in Section 4 are intended to be a more in-depth examination of conditions associated with the performance of each model. As such, they tend to be more related to diagnostic evaluation than to opera-

tional evaluation. The two fractional bias statistics, FB and FS, are shown for the following three data subsets for each model: (1) the six monitoring stations, (2) four stability categories, and (3) three wind speed categories. Also fractional bias of the average is shown as a function of station downwind distance and hour of day. The curves shown on the downwind distance plots are derived using a least squares smoothing algorithm.[10] The results shown in Section 4 reveal patterns of model performance that should be of interest to both the regulatory community and to those attempting to understand and improve models.

The data used in the graphical presentations in both Section 3 and Section 4 are divided into two major groups - (1) the high 25 concentrations, unpaired in time and/or space and (2) all concentrations paired in space and time. The high 25 concentration grouping was selected since these values are of most interest from a regulatory perspective; the all concentration grouping provides a measure of model performance over all measured events of interest to model developers. The number of values comprising the high 25 concentration grouping is always constant since each data subset analyzed (e.g., stable conditions) contains at least 25 values. The number of pairs of values comprising the all concentration group depends on averaging period and data subgrouping. For example, the number of 24-hour values typically consists of hundreds of data pairs while the number of 1-hour values may exceed 10,000.

Since one major purpose of these comparisons is to distinguish between the models' performance, the question of statistically significant difference arises. Because of the complex nature of the comparisons being made

(i.e., ratios involving both observed and predicted values), formal statistical tests of significance are not readily available for the statistics plotted. However, preliminary analyses performed earlier using the Clifty Creek data base resulted in confidence intervals for the difference in the fractional bias of the average for two models.[7] The analysis was performed using the bootstrap procedure in which the standard error was calculated for the difference applied to the high 25 concentration data grouping. That analysis showed that differences in fractional biases for the average are statistically significant at approximately the 5% level if they are separated by more than 0.3 units. This value (0.3 units) was derived using only the highest 25 values, unpaired in space or time, and therefore should be considered as only a rough approximation, especially for data subgroups involving diagnostic related graphs, i.e., those shown in Section 4.

Section 3

SUMMARY OF MODEL PERFORMANCE

In this section, the operational performance of the four models as applied to the Clifty Creek data base is compared using the graphical formats discussed previously. The goal is to characterize the performance of the models in terms of the two fractional bias statistics (FB and FS) and to compare the information conveyed by the FB vs FS plots with that conveyed by (1) the Q-Q plots and (2) the frequency distribution plots.

3.1 RESULTS FOR HIGH 25 DATA

Figure 5 is comprised of six plots corresponding to the averaging periods of 1, 3, and 24 hours for the two years 1975 and 1976. The data used to generate these plots consists of the 25 highest observed and predicted values, unpaired in space or time.

For 1-hour averages, MPTER and TEM are relatively unbiased. MPTER slightly overpredicts the average observed value in each year but slightly underpredicts the scatter in 1975. TEM tends to be unbiased for the average of the high 25 values but overpredicts the scatter for both years. Both MPSDM and PPSP tend to overpredict the average and scatter in excess of a factor-of-two. As the averaging period increases, the models shift directionally toward less overprediction. For the 24-hour averaging period, PPSP continues to significantly overpredict while TEM tends toward significant underpredictions. The other two models, MPSDM and MPTER, exhibit the least overall bias for 24-hour averages. Since a difference of approximately 0.3 units between two fractional bias statistics is assumed to be statistically significant, PPSP has a bias of the average for the 25 highest

values that is clearly different from any of the other three models. While differences among the other three models approach statistical significance for some averaging periods, MPSDM, MPTER and TEM are much closer in performance as a group than PPSP.

To further illustrate the difference in performance among the four models, Figure 6 (see reference 7) is presented. In Figure 6, the probable range of outcomes for each of the four models is shown for 24-hour averages in 1975. The elliptically shaped clusters are the results of 200 samples using the bootstrap procedure.[11] The results clearly indicate the significance of the PPSP overpredictions and also demonstrate the overlap between MPTER and the other two models, MPSDM and TEM. Because computations are relatively expensive, the bootstrap is performed and illustrated only for this particular data group. Nevertheless, the reader should obtain some sense of the uncertainty associated with any given plot involving the FB and FS statistics.

A pattern in the FB vs FS plots (Figure 5) and subsequent plots is worth noting. Namely, there is a tendency for underpredictions in scatter to be associated with underprediction in the average (upper right quadrant) and, similarly, overprediction in scatter to be associated with overpredictions of the average (lower left quadrant). This pattern is consistent with any tendency for a model to over or underpredict the observed value by a constant ratio. If a model overpredicts by a constant multiple of the observed concentration, both the average and the scatter will also be overpredicted by the same multiple. The result is that a model that overpredicts by a constant factor will plot in the lower quadrant near a diagonal line through the origin that defines equal values for the fractional bias of the

average and the standard deviation. Conversely a model that underpredicts by a constant factor will plot in the upper quadrant very near the same diagonal line. The extent to which a model plots some distance away from this diagonal line is a measure of the tendency for the model to behave counter to the hypothesis of constant over or underprediction.

For example, in Figure 5, PPSP tends to overpredict both bias statistics by nearly an equal degree; this is not inconsistent with an hypothesis of overprediction by a constant factor. However, a comparison between 1975 and 1976 for 1-hour values shows that PPSP falls slightly above the diagonal for 1975 and somewhat below the diagonal for 1976. Thus while PPSP clearly overpredicts both bias statistics for each year, the two years differ in that overprediction of the scatter in 1975 is less than overprediction of the average (FS = -1.0 vs FB = -1.3), while the opposite is true in 1976 (FS = -1.8 vs FB = -1.4). The same finding is generally applicable to the other models and averaging periods with no clear exceptions.

The six Q-Q plots shown in Figure 7 are created from the same data used to generate Figure 5. The Q-Q plots permit a visual inspection of both the magnitude and rate of change of predicted concentrations with increasing observed values; whereas, the fractional bias plots summarize fractional bias of the average and scatter irrespective of concentration magnitude. Several features are worth noting about these plots and how they compare with the previous figure. First, the bias exhibited by PPSP is more obvious. Second, MPTER would appear to be relatively unbiased for the entire range of observed 25 highest values since the data points for MPTER are consistently close to the line of equal observed and predicted values for each averaging period for both years of data. MPSDM and TEM

also tend to be accurate within a factor of two for some of the averaging periods. However, TEM clearly underpredicts for 24-hour averages for a full range of the high 25 values.

The slopes of the Q-Q plots convey information that is related to that contained in the statistics shown above in Figure 5. For example, the slopes of the Q-Q plots of 1-hour averages for PPSP are somewhat different between 1975 and 1976. For 1975, the degree of overprediction, measured by the distance between the predicted points and the diagonal line of equal observed and predicted values, tends to decrease as the observed concentration increases (i.e. slope of data points is slightly less than one); the opposite trend is evident in 1976. Since the Q-Q plots are scaled logarithmically, a slope of less than one indicates that relative scatter in predicted values is less than the relative scatter in observed values. This explains in Figure 5 why PPSP plots above the diagonal line for 1975 and below the diagonal line in 1976.

3.2 RESULTS FOR ALL DATA

The plots shown in Figures 8 and 9 are companion plots to those shown in Figures 5 and 7. The difference is the data used to develop Figures 8 and 9 represents all concentrations paired in time and space -- not just the high 25 values.

The same basic trends as shown in Figures 5 and 7 exist, i.e., the models tend toward greater underprediction as averaging period increases. They differ however in that Figure 8 clearly indicates that the models as a group tend towards larger underpredictions when all data are used than is the case for the 25 highest values only. In fact, MPTER, MPSDM and TEM

systematically underpredict the average for all averaging periods. The bias towards underprediction is least for MPSDM and greatest for TEM which underpredicts generally by a factor-of-two or more. The exception appears to be PPSP for which overpredictions exceeding a factor-of-two are still the rule for all averaging periods.

Figure 9 presents cumulative frequency distributions for the all concentration data group. None of the distributions approach a straight line, indicating that neither the predicted values nor the observed data approximate a log-normal density function. Again PPSP strongly overpredicts the upper percentiles but tends to underpredict the concentrations below approximately 30 to 50 $\mu g/m^3$. The other three models fit moderately well over the upper 5 percent of the data; however, underpredictions by TEM are evident especially for 24-hour averages.

3.3 OPERATIONAL CONCLUSIONS

Bias, Q-Q and frequency distribution plots are used to graphically assess the ability of four models to accurately reproduce observed concentrations for several averaging periods. From those graphical presentations the following conclusions are drawn:

1. The various graphical presentations are consistent in what they show about model performance; however, each contains unique information which supplements the others; the bias plots appear to have the greatest flexibility and effectively summarize information for further use in the diagnostic evaluation presented in Section 4;

2. All models tend toward less overprediction and/or greater underprediction for longer averaging periods;

3. MPTER shows the least bias for the full range of concentrations for all averaging periods; PPSP shows consistent bias to overpredict concentrations; TEM shows consistent bias to underpredict concentrations for the highest 24 hour average concentrations; MPSDM shows variable performance for the 25 highest concentrations, but the least overall bias when all concentrations are paired in space and time.

4. The relative performance among the four models is strikingly consistent for each of the two years; however, subtle differences between years are detectable. For example, performance results for 1976 compared to 1975 tend toward greater underprediction of the average and greater overprediction of the scatter as evidenced by values that plotted below the diagonal line in the 1976 bias plots and slopes greater than unity in the 1976 Q-Q plots.

SECTION 4

MODEL PERFORMANCE BY DATA SUBSETS

The information presented in this section is intended to provide a preliminary framework for diagnostic-related evaluations using some of the graphical formats presented earlier. This section presents the fractional bias of the average for various data subcategories in order to highlight performance trends by downwind distance, meteorological categories, and time of day. To minimize the volume of information shown, only results for 1-hour averages for 1975 are presented. Results for 1976 (not shown here) indicate basically the same trends and patterns. Also, to assist the reader in understanding model performance by receptor and downwind distance, Figure 10 taken from the paper by Irwin and Smith[9] is included. This figure illustrates the nature of the terrain between the source and the monitoring stations. It shows the terrain height for each station and the terrain cross section between each station and the source. It should be noted that stations 1 and 4 are the most distant on elevated terrain, station 5 is the closest on elevated terrain, stations 2 and 3 are at an intermediate distance on elevated terrain, while station 6 is at plant grade.

Figures 11-13 present details for each station, each stability category, and each wind speed category; each uses the 25 highest concentrations unpaired in space or time. Figures 14-16 are companion figures that show the same information, except that all concentrations paired in space and time are used. Each of the 6 figures consists of four plots corresponding to the four models. The symbols plotted correspond to the station numbers (Figures 11 and 14), stability categories (Figures 12 and 15) and wind speed categories (Figures 13 and 16).

Diagnostic graphs shown in Figures 17-19 depict fractional bias of the average as a function of station-source distance for all meteorological events combined, and separately, for each of the various meteorological categories. Figures 20-22 are similar plots except that all concentration data paired in space and time are used.

Finally, Figures 23-26 present fractional bias as a function of hour of day for all stations combined and each station separately. The 25 highest concentrations unpaired in space or time are shown in Figures 23 and 24; all concentrations paired in space and time are shown in Figures 25 and 26.

4.1 RESULTS BY MODEL FOR INDIVIDUAL STATIONS AND METEOROLOGICAL SUBSETS

In Figure 11, TEM appears to exhibit the lowest bias for the monitoring stations as a whole; each station falls within a factor-of-two for both fractional bias statistics. MPTER is comparable with a slightly wider range of performance among stations. It is evident that overpredictions for PPSP exceed a factor-of-two at each of the 6 stations. MPSDM shows a general tendency to overpredict for all stations and by more than a factor-of-two at several. An additional interesting feature is the general consistency in the relative clustering of the stations across models. Concentrations for Station 5 are the most systematically overpredicted of the 6 Stations which may be related to the fact that this station is the closest elevated receptor (see Figure 10). This similarity may be attributed to the fact that all models are Gaussian, and thus do not treat atmospheric transport and dispersion in fundamentally different ways.

In Figure 12, results are shown for the four stability categories which are comprised of the Pasquill-Gifford classes A-G as follows: very

unstable--class A or B, unstable--class C, neutral--class D, and stable--class E, F or G. While some general tendencies are noticeable, each model is somewhat unique with regard to the scatter and placement of the bias statistics. The stable category is generally associated with the greatest underpredictions while the unstable category is associated with the greatest overpredictions. The model's differ however in the degree to which this tendency is true. MPSDM shows the greatest sensitivity to stability category while MPTER and PPSP show the least sensitivity. MPTER appears to be the most accurate of the four models across the four stability categories since only the fractional bias of the standard deviation for stable conditions lies outside of a factor-of-two. TEM appears to perform best overall for unstable conditions.

Figure 13 shows the results for three wind speed categories defined as follows: low -- less than 2.5 mph, medium -- 2.5 to 5.0 mph, and high--greater than 5.0 mph. Again MPTER appears to be the most accurate model since the fractional biases are tightly clustered and well within a factor-of-two accuracy. MPSDM exhibits the greatest sensitivity to wind speed category. TEM shows a significant departure from previously observed patterns between the two fractional bias measures; for the high wind speed category, TEM tends to underpredict the average observed value by a factor-of-two, while it overpredicts the scatter in the observed data by greater than a factor-of-two. This causes TEM to be somewhat removed from the diagonal line of equal fractional bias for the average and scatter (refer to discussion in Section 3).

Figure 14 shows results for each station for the all concentration data category. Compared to Figure 11, a general trend towards less overprediction occurs when all data are used. For MPTER and TEM slight overpredictions become major underpredictions and for MPSDM the major overprediction is significantly reduced. Station 5 shows a noticeable shift to less overprediction when all data are used for MPTER, MPSDM and TEM. There is little change for PPSP at any station in the overall amount of overpredictions.

Figure 15 is the companion to Figure 12 and shows results for each stability category when all data are used. Overprediction appears to be less of a problem except for the more unstable categories where the predicted scatter significantly exceeds the scatter in the observed concentrations for each of the models. Comparison of the two figures (Figure 12 and 15) reveals that except for PPSP, the neutral category appears to shift more dramatically towards underprediction than for the other categories. The range between stabilities remains large with stable and neutral categories associated with underpredictions and unstable conditions associated with overpredictions.

Figure 16 completes the meteorological subset comparisons for wind speed categories with the all concentration data set. MPTER again appears to perform best since the fractional bias for each of the wind speed categories indicates performance that is within a factor-of-two. The trend for less overprediction when all data are used, is evident. MPTER slightly underpredicts averages for each category while it slightly overpredicts the scatter.

4.2 STATION DISTANCE PERFORMANCE PATTERNS

A series of similar bias plots are presented in Figures 17, 18, and 19. The fractional bias of the average of the high 25 values is plotted as a function of the distance between the source and each of the six stations. Figure 17 shows results for all meteorological subsets combined. Figure 18 shows similar results for the four stability categories while Figure 19 shows the results for low, medium and high wind speed categories respectively. The curves for the four models are best fit lines obtained using a least squares smoothing algorithm[10]. Some interesting patterns emerge from these plots. One, there seems to be a general tendency for the fractional bias to be larger in magnitude at the closer stations and smaller at the more remote stations. Two, each model exhibits a reasonably well defined distance trend that is strongly dependent on the meteorological subset represented.

Considering Figure 17 in more detail, it appears that TEM and MPTER show the least sensitivity in performance with distance and also show the least overall bias. At the other extreme, PPSP shows large overprediction at the closest stations, but this decreases for the more distant stations.

Examination of the stability plots (Figure 18) reveals pronounced trends for the four models. For very unstable conditions, all four models show a tendency for decreasing overprediction as distance between source and receptor increases. All four models overpredict significantly at the closest station; at the most remote station TEM is essentially unbiased, while the other three models continue to show slight overprediction. As the stability increases, this pattern continues for some of the models

while others show a reversal of this trend. For example both TEM and MPTER show decreasing underprediction (rather than increasing) with distance for unstable and neutral categories, while the previous pattern holds for PPSP and MPSDM. For stable conditions, only PPSP continues to exhibit the same pattern as evident for very unstable conditions, i.e. a tendency for decreasing overprediction with increasing distance.

For the wind speed category plots (Figures 19), the contrast between the low and high speed categories is evident. Although patterns are different among the models, the general tendency is for low wind speeds to trend toward less overprediction with distance while for the high wind speeds the tendency is reversed, i.e. decreasing underprediction as distance increases. Interestingly, results for PPSP, which generally overpredicts, show a consistent unbiased result for all source/stations separation distances for the high wind speed category.

In most of the distance plots, concentrations for Station 6 at about 5 miles appear to be underpredicted relative to the general trend indicated by the smooth curve. From Figure 10, it can be seen that station 6 is more than 300 feet lower than the other stations; this appears to have an affect on the relative performance of the models at that location.

Figures 20-22 present the same type of information shown in Figures 17-19 except that all data paired in space and time are used for each hour. The basic patterns described above are the same. Also the general tendency for the all concentration group to be associated with greater underprediction is obvious. The major difference is reflected in somewhat flatter curves for some of the models, especially for MPSDM.

4.3 DIURNAL PERFORMANCE PATTERNS

Figures 23 and 24 present diurnal patterns of model performance in which the fractional bias for averages is plotted as a function of hour of the day using the high 25 concentrations for each hour. Figure 23 shows the results for all stations combined, while Figure 24 shows the results for each individual station. Overall a rather striking pattern emerges for each model; the pattern consists of pronounced underprediction during both the early morning and the evening hours, and pronounced overprediction during the midday hours. PPSP exhibits the greatest difference in performance with fractional bias ranging from values near +2.0 in the early morning and evening to near -2.0 in the late morning and the late afternoon hours. For PPSP, there is also a noticable trend for fractional bias to improve (less overprediction) around midday followed by the decline in late afternoon which creates a "W" shaped pattern for the day. MPTER and MPSDM appear to have the most consistent performance as indicated by the relatively small range in the bias across the day. TEM underpredicts very significantly during the morning and evening but is relatively unbiased for the midday hours.

There are differences in the diurnal patterns among the individual stations that warrant attention, especially for MPSDM and MPTER. For the two most distant stations (Stations 1 and 4), the range in fractional bias is narrow and consistently close to zero. The fractional bias for MPSDM and MPTER at these two stations does not fall below -0.6 nor exceed approximately 0.8, indicating a level of performance that is within nearly a factor-of-two for every hour of the day. This consistency contrasts sharply with performance at the station closest to the stack (Station 5), where MPTER signifi-

cantly underpredicts for most hours while MPSDM swings markedly from large overpredictions through large underpredictions. Diurnal patterns at the station located at the lowest terrain (Station 6) indicates that MPSDM has a relatively small bias across the day compared with that for the other three models. PPSP exhibits the same basic pattern at each station, i.e., a tendency to underpredict early morning and late evening values while severely overpredicting mid-day hourly values. TEM also exhibits the same pattern at each station with mid-day hourly predictions being essentially unbiased.

Figures 25 and 26 present the same plots using all concentration data for each hour of the day. Basically the patterns are the same with a tendency for overprediction (or less underprediction) during the midday hours and underprediction otherwise. The range in the fractional bias is similar between the two data groups, except for PPSP at Station 1 where the degree of overprediction is not as severe for the all concentration data group.

4.4 DIAGNOSTIC CONCLUSIONS

Various forms of fractional bias plots are used to better diagnose model performance for various subsets of information including stability class, wind speed category, downwind distance, and time of day. From these graphical presentations, the following conclusions are drawn:

1. Considerably more detail is provided as to those factors contributing to results shown in Section 3 for the operational evaluation;

2. There appears to be a clear variation in accuracy by stability class with the models tending to overpredict for unstable conditions and

underpredict for stable conditions; TEM shows the least overprediction for unstable conditions, while MPTER appears to show the least overall bias;

3. For wind speed categories there is a wide disparity in model performance for all models, except for MPTER which shows low overall bias across the three categories; generally the least overprediction occurs for the high wind speed category;

4. Variations in performance among the stations are clearly evident with the models showing the least bias for the most distant stations; underpredictions and overpredictions appear to be accentuated for stations closer to the source; smaller overpredictions or greater underpredictions are evident for the one station located at plant grade.

5. There are distinct differences in how all the models perform for time of day with all tending to underestimate in the noctural hours and to overestimate during hours of strong solar radiation; this is undoubtedly associated with parallel biases shown for stability classes; the most pronounced differences occur for PPSP and TEM; somewhat smaller differences occur for MPTER and MPSDM, but there are important variations from station-to-station.

SECTION 5

SUMMARY AND CONCLUSIONS

A simple graphical format has been used to present summaries of operational model performance using two statistics -- (1) the fractional bias of the average and (2) the fractional bias of the standard deviation. The format was used to display and compare the performance of four rural models previously evaluated for Clifty Creek. The information was conveyed in a convenient and readily understandable manner especially suitable for officials concerned with air quality regulation and management. Additional information provided in supplementary Q-Q and frequency distribution plots was shown to be related to the fractional bias statistics but supplied greater detail regarding the magnitude of observed and predicted discrepancies.

Several graphical formats were presented that are of value in diagnosing model performance. The fractional bias was displayed for each station, wind speed, and stability class making semi-quantitative but visual analyses possible. These analyses revealed conditions associated with consistently unbiased performance, and conversely, conditions associated with inconsistent or biased performance. Similar plots showing fractional bias as a function of hour of day and downwind distance proved valuable in examining the magnitude and consistency of model bias both diurnally and across terrain between the source and monitoring stations. The graphical formats and data subsets presented can be used as a beginning for development of a framework for standardization of diagnostic performance evaluations.

From these graphical presentations it was possible to obtain a clearer understanding of factors that contribute to overestimates and underestimates

by the models. The diagnostic tools used here are intended to provide a standardized, objective approach. A much more careful and thorough event-specific analysis of each model is necessary to fully understand their faults and to provide a basis for research into improving the models. Nevertheless, from the information presented here it is clear that the interrelationship between downwind distance, stability class, and time of day play a dominant role in biases exhibited by these models and should receive careful attention in efforts to improve the models.* It would also seem that qualitatively MPTER exhibited the least overall bias of the four models for the graphical presentations considered to represent the Clifty Creek data base.

In conclusion, further testing of these techniques seems warranted to develop additional graphical formats and/or data groupings and for application to other data bases.

* It is recognized that the interrelationship between plume rise and mixing height, which can also affect the biases considered here, could not be analyzed due to limitations of the Clifty Creek data base.

REFERENCES

1. R. J. Londergan, D. H. Minott, D. J. Wackter, T. Kincaid and D. Bonitata, "Evaluation of Rural Air Quality Simulation Models," EPA-450/4-83-003, October 1982.

2. Cox, W. M. and Gerald K. Moss, "Evaluation of Rural Air Quality Simulation Models, Addendum A: Muskingum River Data Base," EPA-450/4-83-003a, June 1985.

3. D. G. Fox, "Judging Air Quality Model Performance," Bull. Amer. Meteor. Soc. 62(5):599 (1981).

4. R. J. Londergan, D. H. Minott, D. J. Wackter and R. R. Fizz, "Evaluation of Urban Air Quality Simulation Models," EPA-450/4-83-020, July 1983.

5. R. J. Londergan and D. J. Wackter, "Evaluation of Complex Terrain Air Quality Simulation Models," EPA-450/4-84-017, June 1984.

6. J. A. Tikvart and W. M. Cox, "EPA's Model Evaluation Program," Paper Presented at the Fourth Joint Conference on Applications of Air Pollution Meteorology, Portland, OR, October 1984.

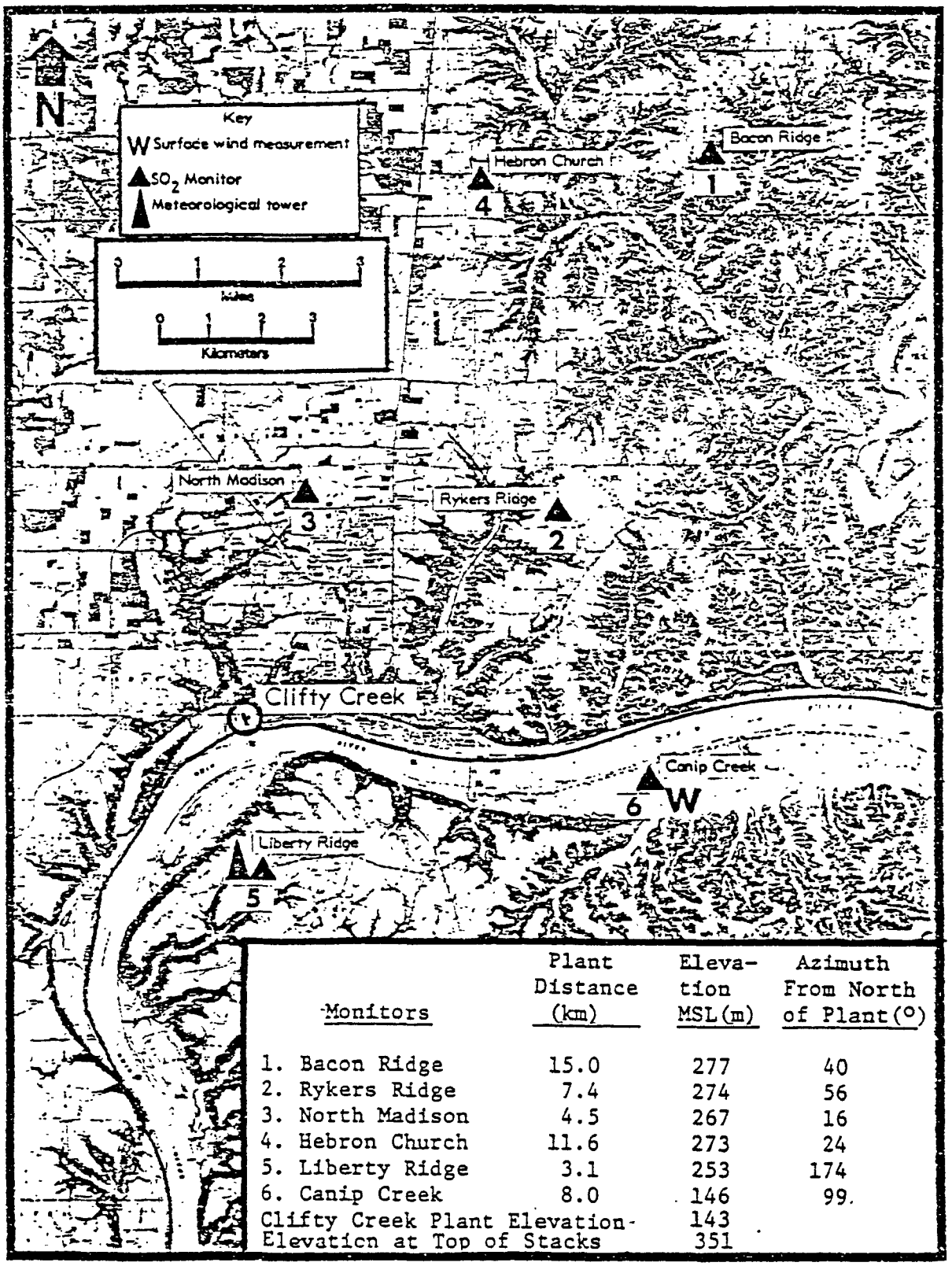
7. W. M. Cox, J. A. Tikvart, "Assessing the Performance Level of Air Quality Models", Paper Presented at the 15th International Technical Meeting On Air Pollution and Its Application, NATO/CCMS Conference, St. Louis, MO, April 1985.

8. W. M. Cox, J. A. Tikvart and J. L. Pearson, "Preliminary Conclusions from EPA's Model Evaluation Program," Paper 85-24A.4 Presented at the 78th APCA Annual Meeting, Detroit, MI, June 1985.

9. J. S. Irwin and M. E. Smith, "Potentially Useful Additions to the Rural Model Performance Evaluation," Bull. Amer. Meteor. Soc. 65(6):559(1984).

10. TELL-A-GRAF Users Manual, Version 5.0 Published by Integrated Software Systems Corporation, 1984.

11. B. Efron and G. Gong "A Leisurely Look at the Bootstrap, the Jacknife, and Cross-Validation," The American Statistician 37(1):36(1983).



Monitors	Plant Distance (km)	Elevation MSL(m)	Azimuth From North of Plant(°)
1. Bacon Ridge	15.0	277	40
2. Rykers Ridge	7.4	274	56
3. North Madison	4.5	267	16
4. Hebron Church	11.6	273	24
5. Liberty Ridge	3.1	253	174
6. Canip Creek	8.0	146	99
Clifty Creek Plant Elevation		143	
Elevation at Top of Stacks		351	

Figure 1. Field Monitoring Network Near The Clifty Creek Power Plant

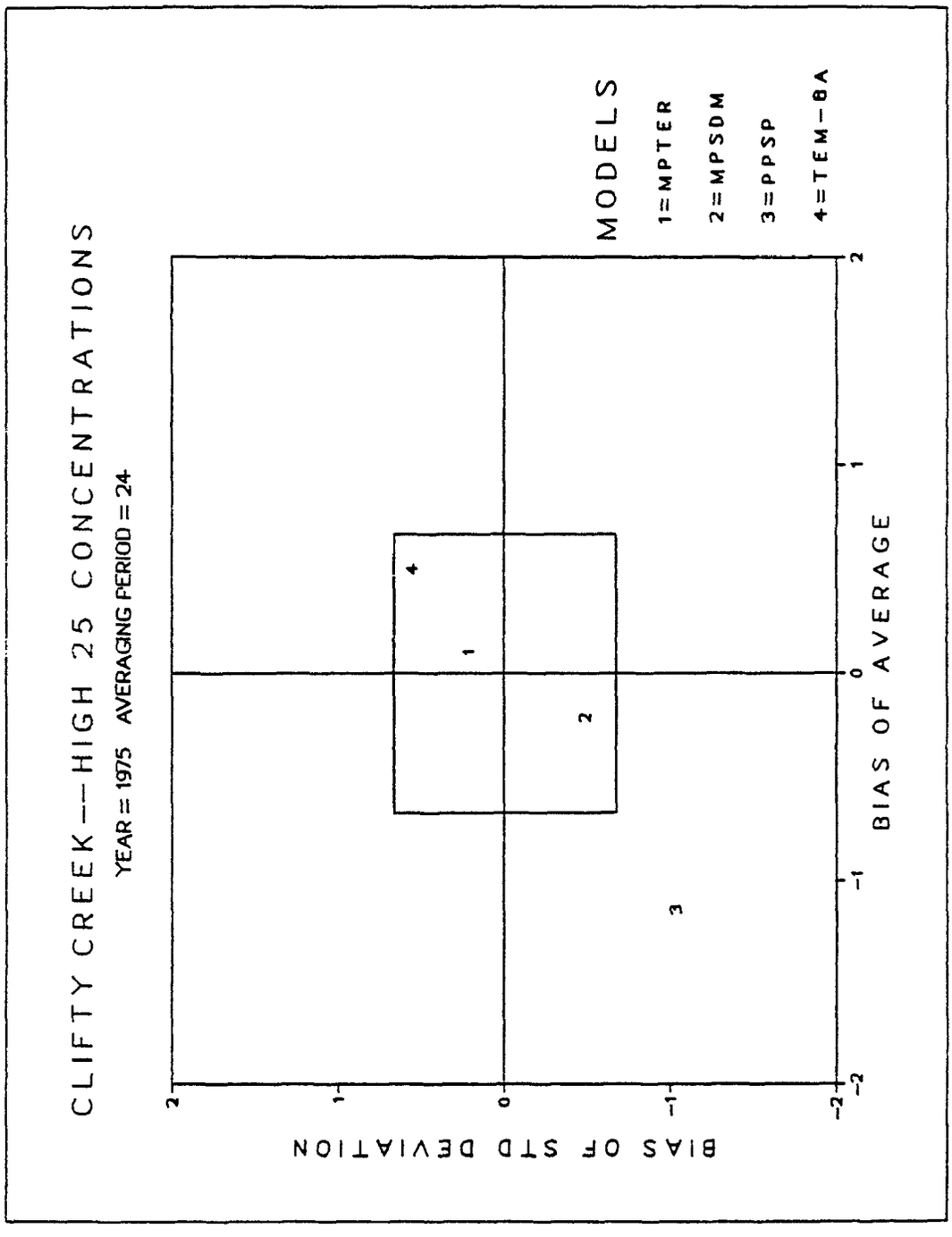


Figure 2. Example Fractional Bias Plot

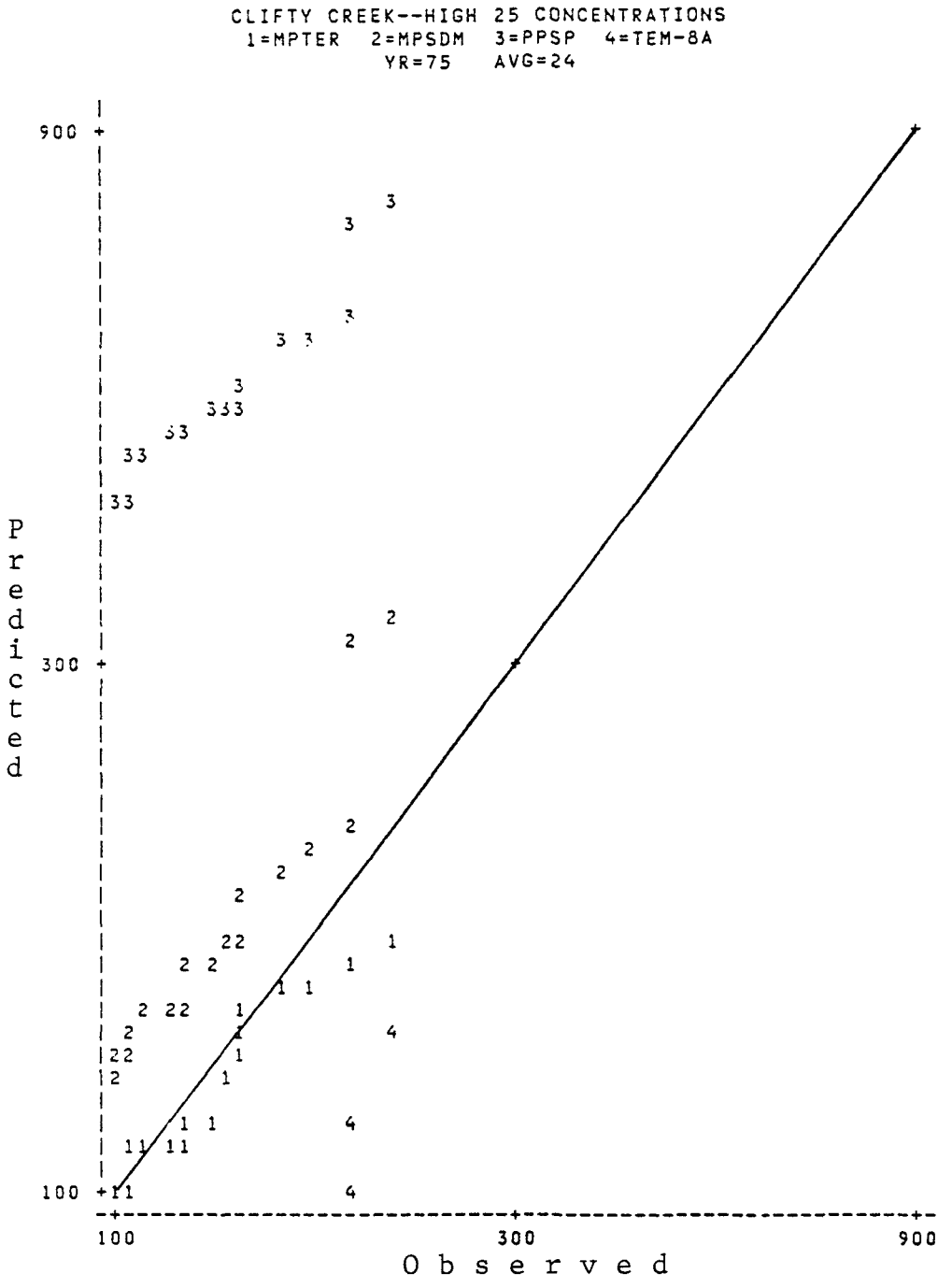


Figure 3. Example Quantile-Quantile Plot

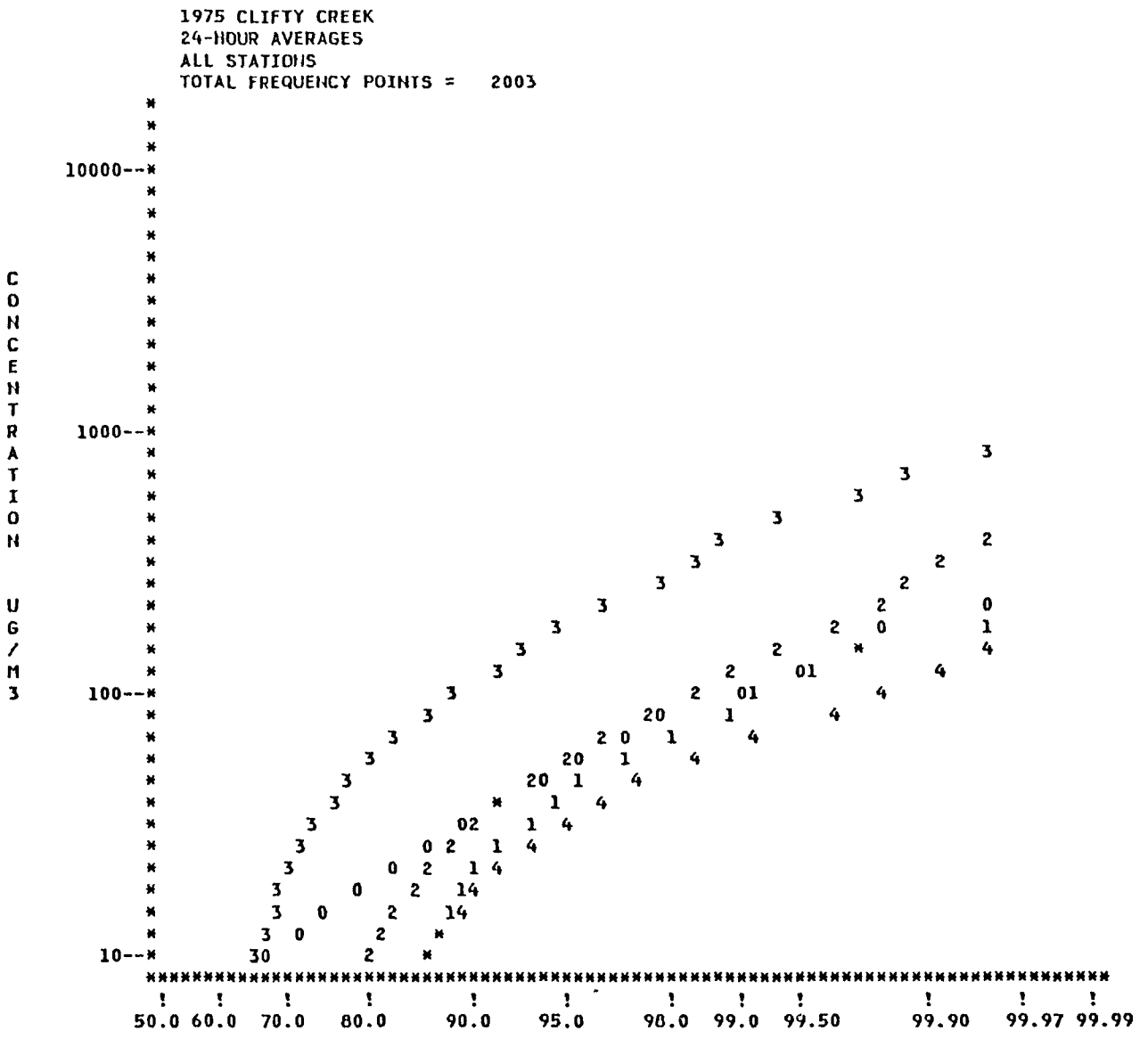


Figure 4. Example Cumulative Frequency Distribution Plot

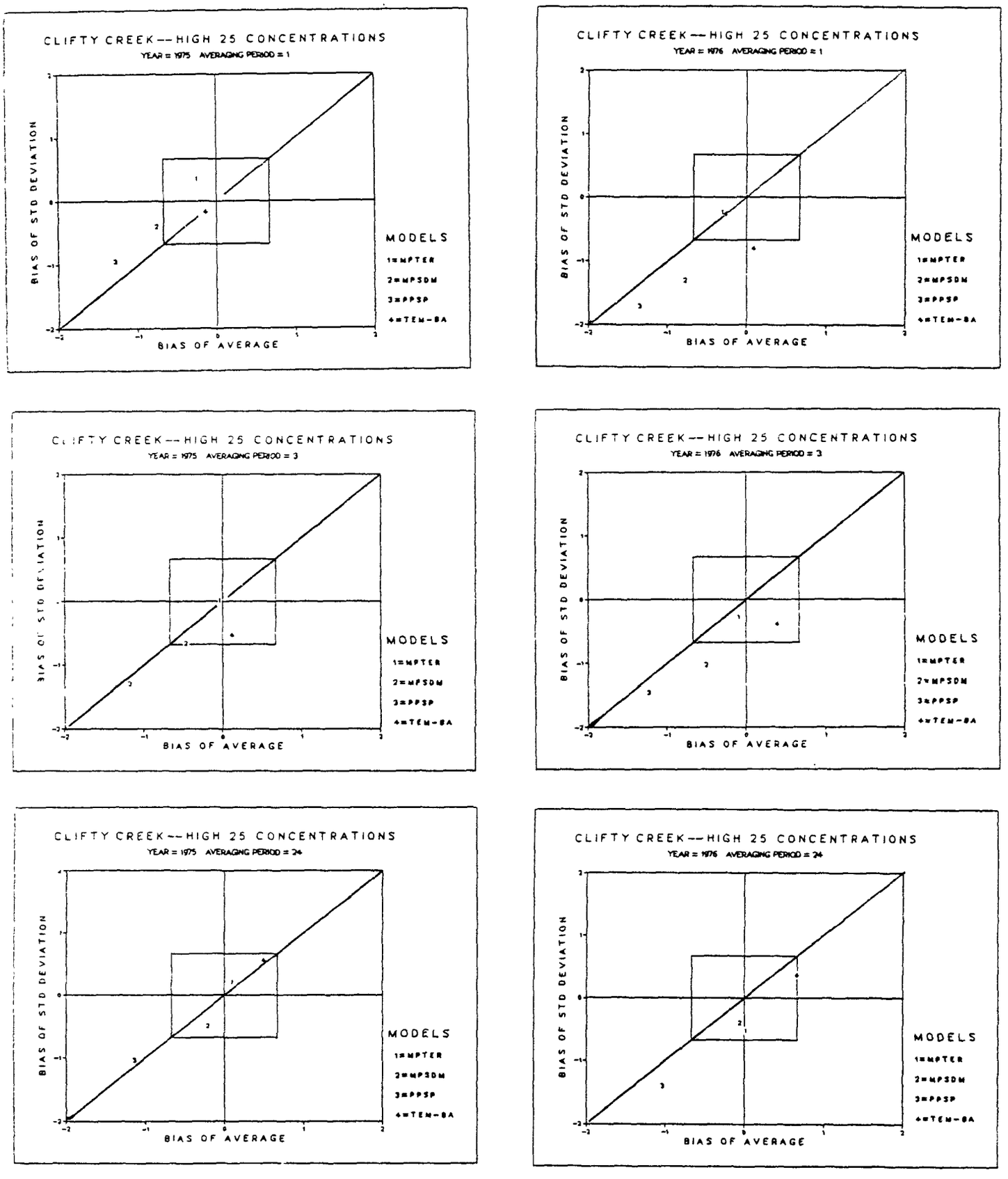


Figure 5. Fractional Bias Plot By Year And Averaging Period Using High 25 Values

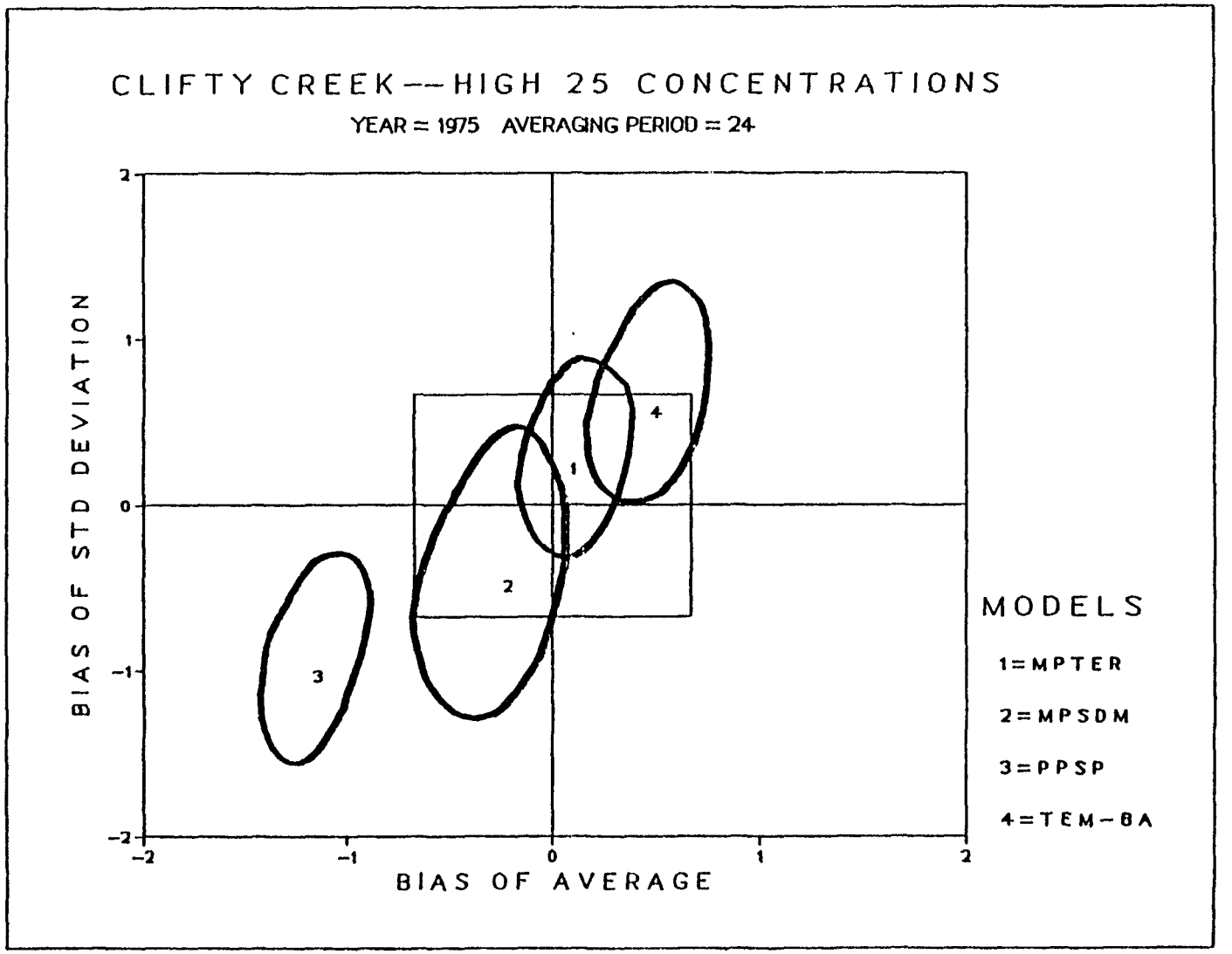


Figure 6. Distribution of 200 Bootstrap Samples: Fractional Bias Plot For 24-Hour Averages Using High 25 Values

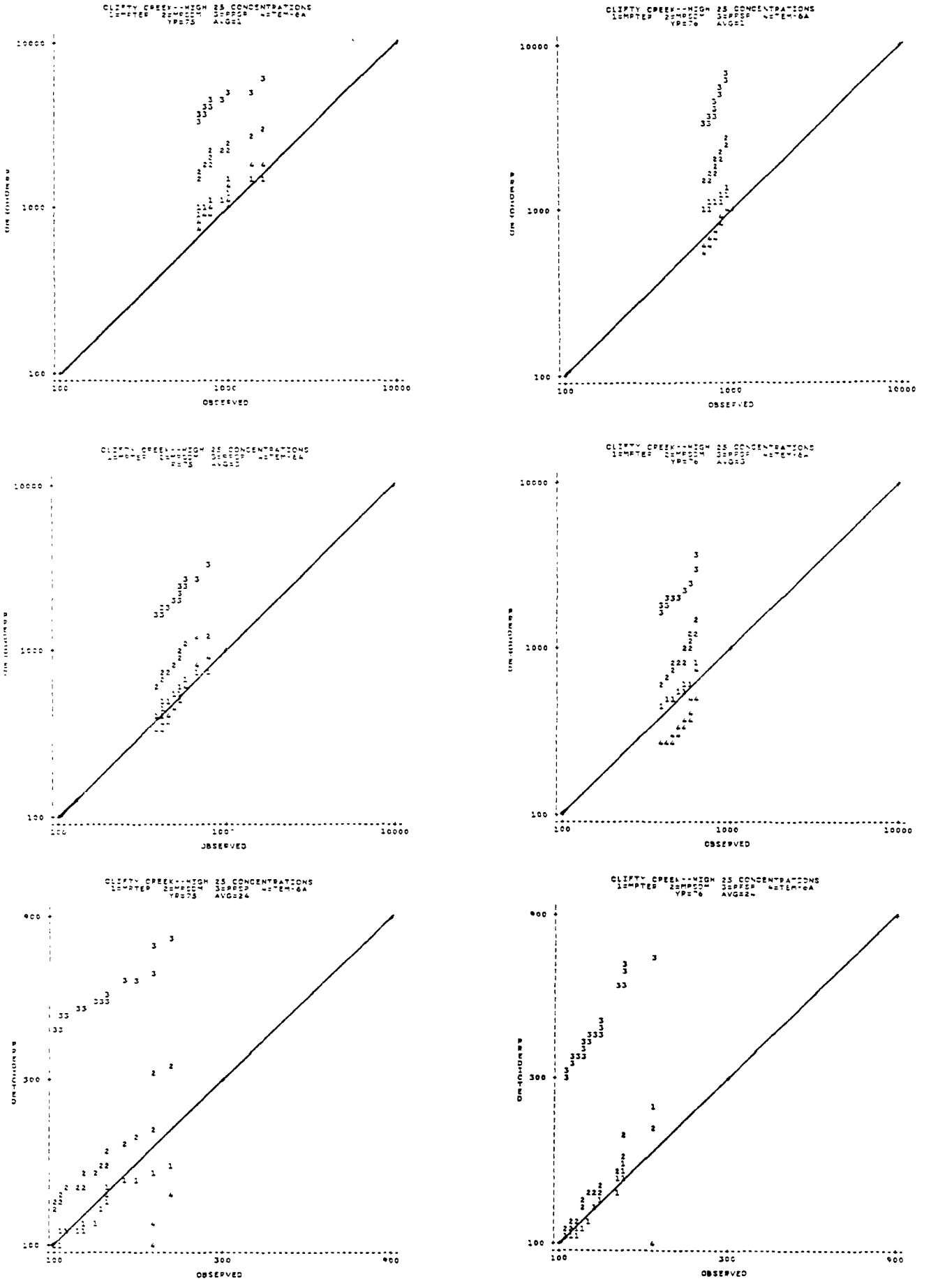
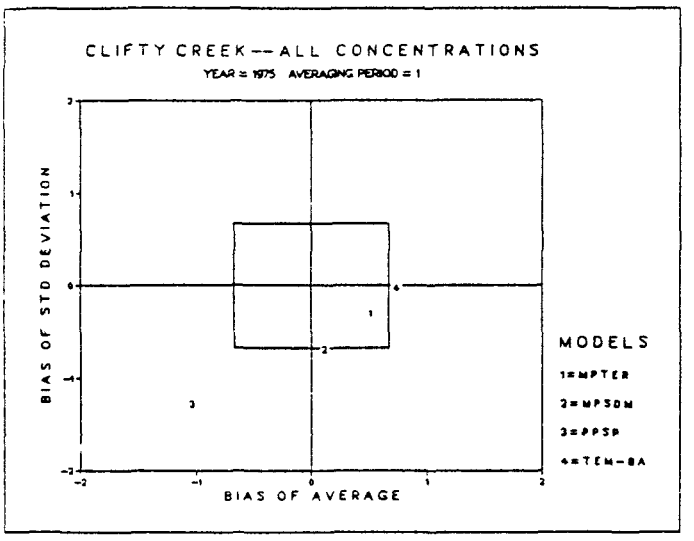
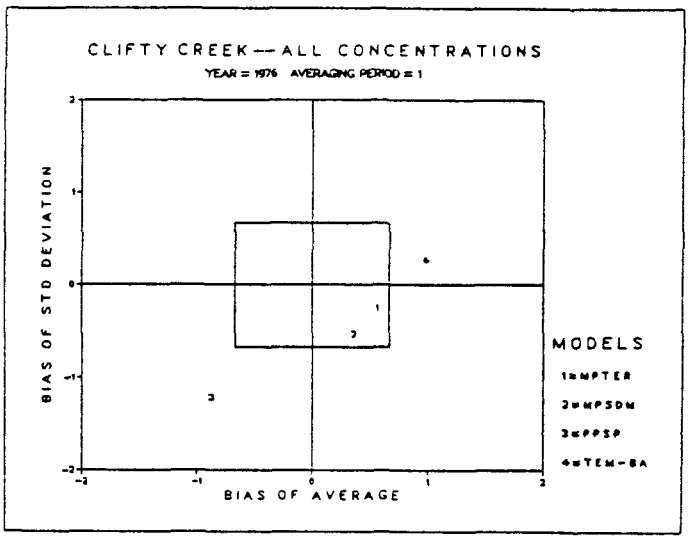
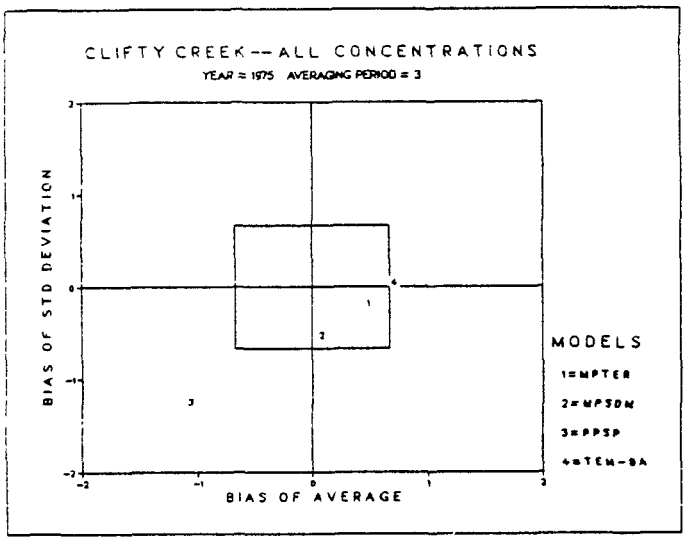
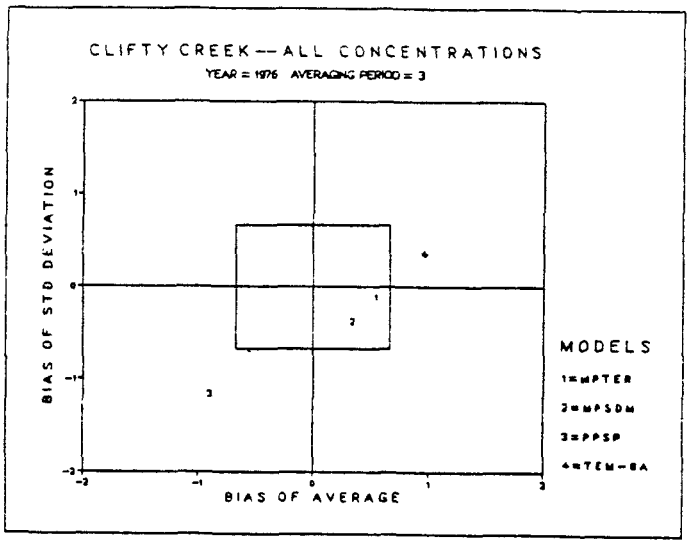


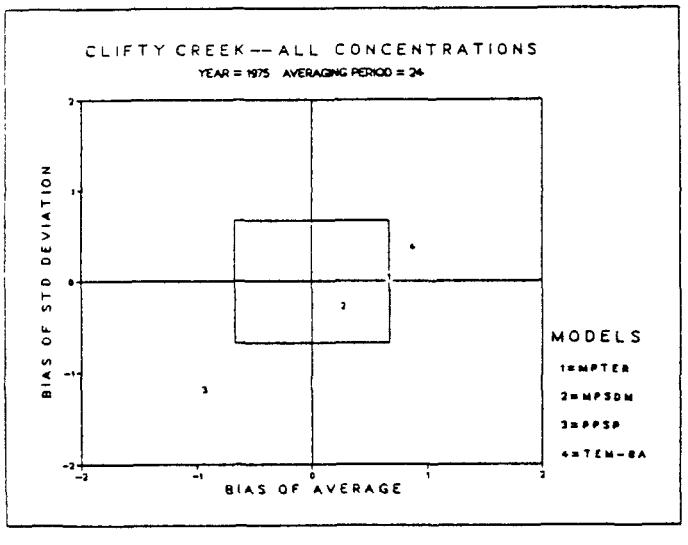
Figure 7. Quantile-Quantile Plot By Year And Averaging Period Using High 25 Values











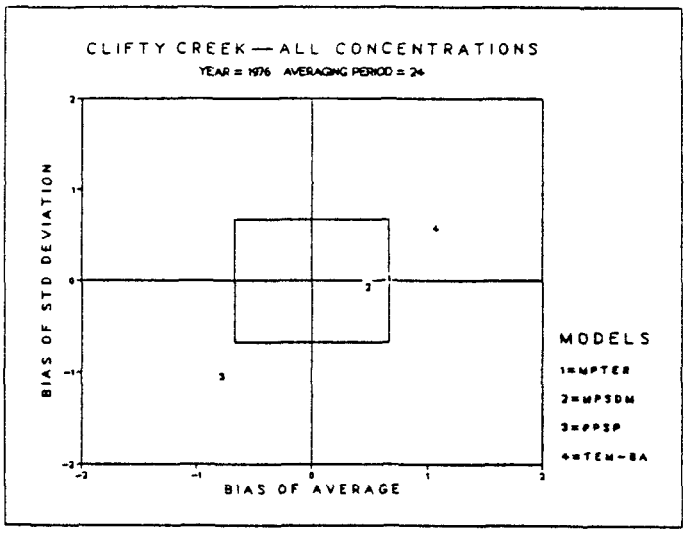


Figure 8. Fractional Bias Plot By Year And Averaging Period Using All Paired Values

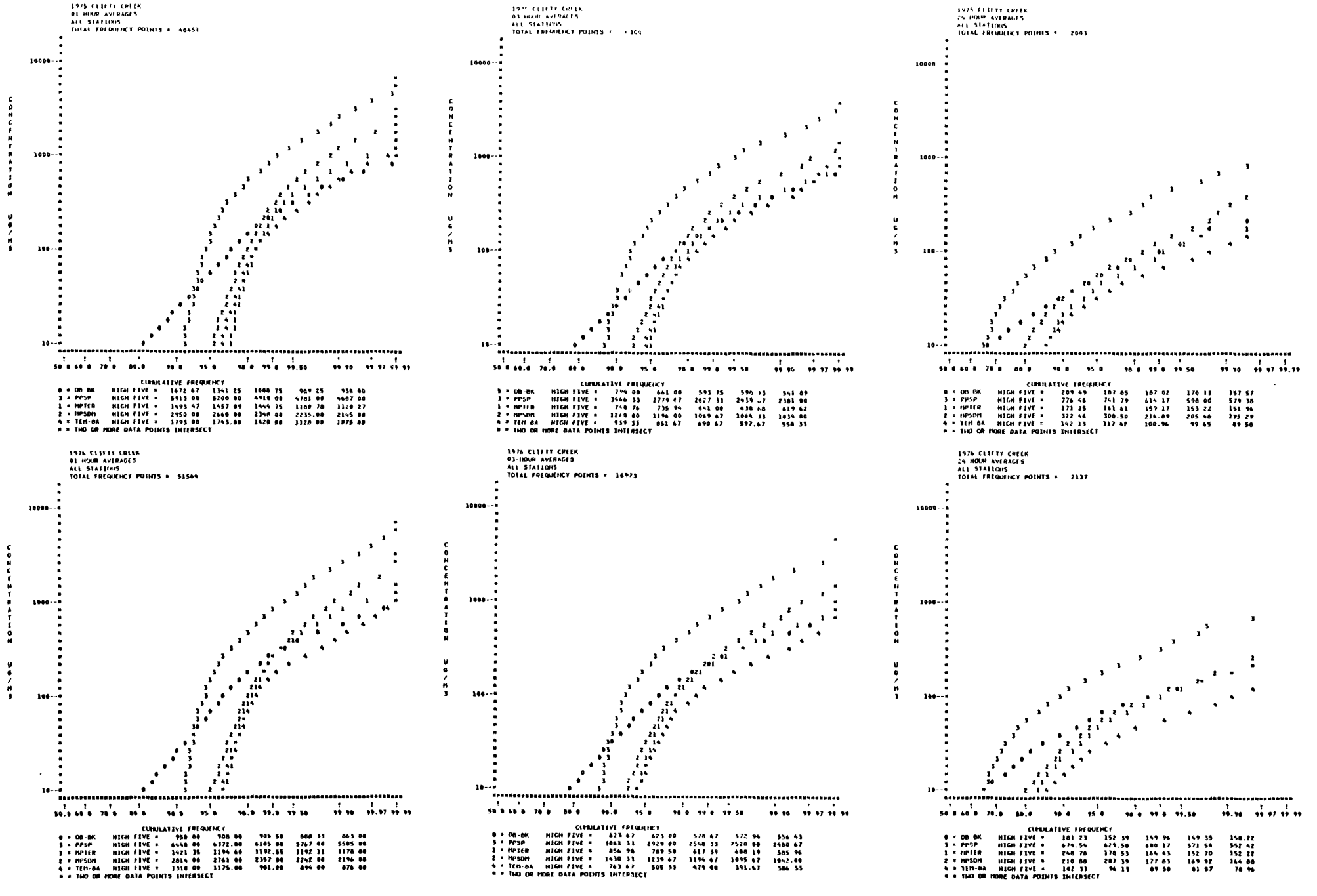


Figure 9. Cumulative Frequency Distributions By Year And Averaging Period Using All Paired Values

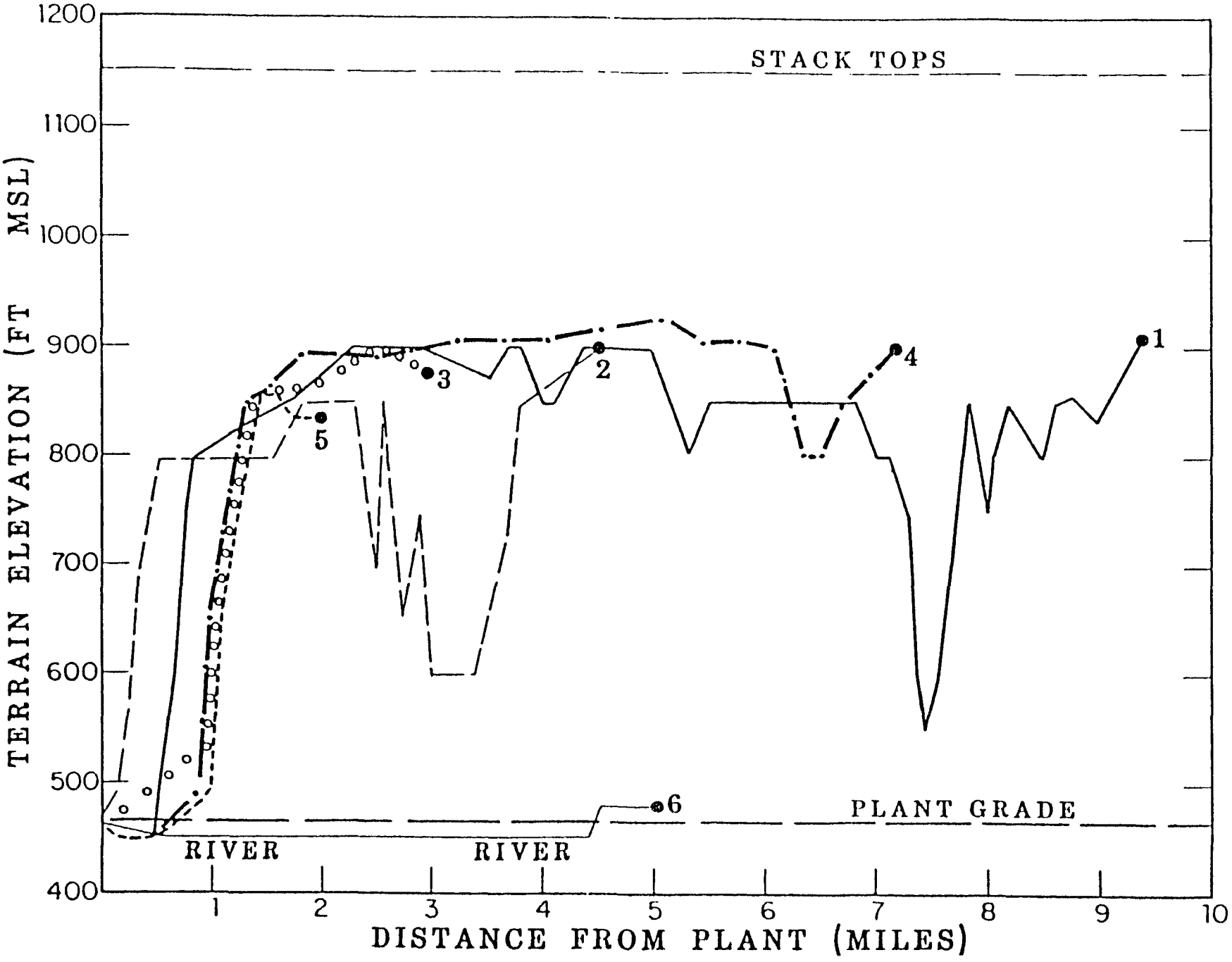
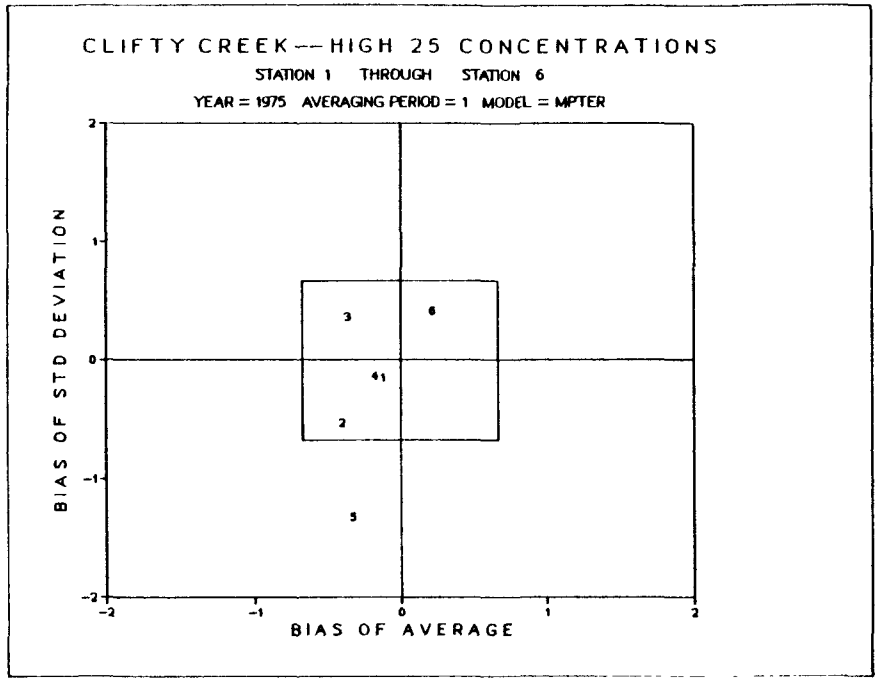
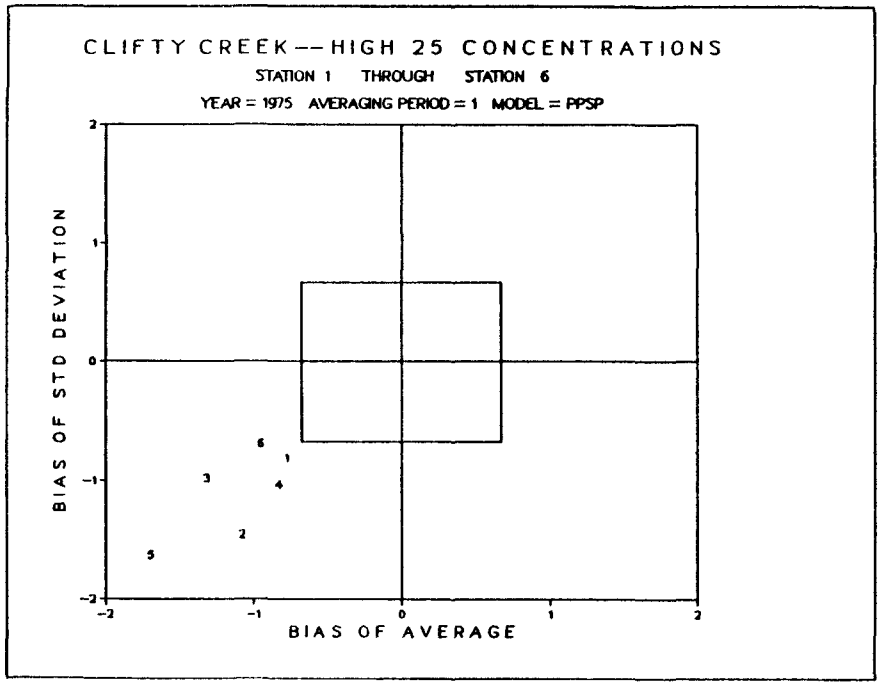
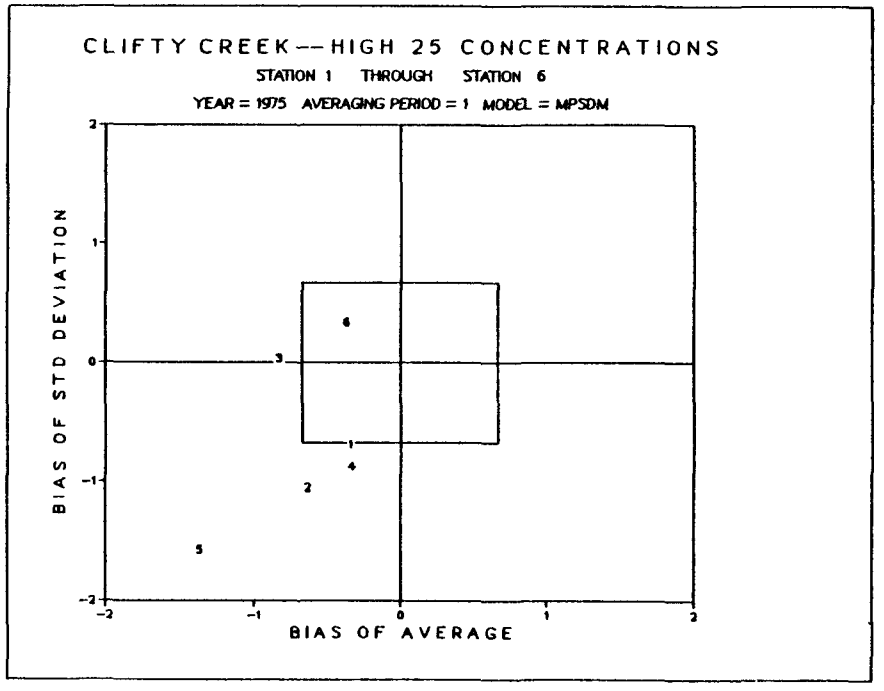


Figure 10. Terrain Profiles Between Clifty Creek Plant and Monitoring Stations







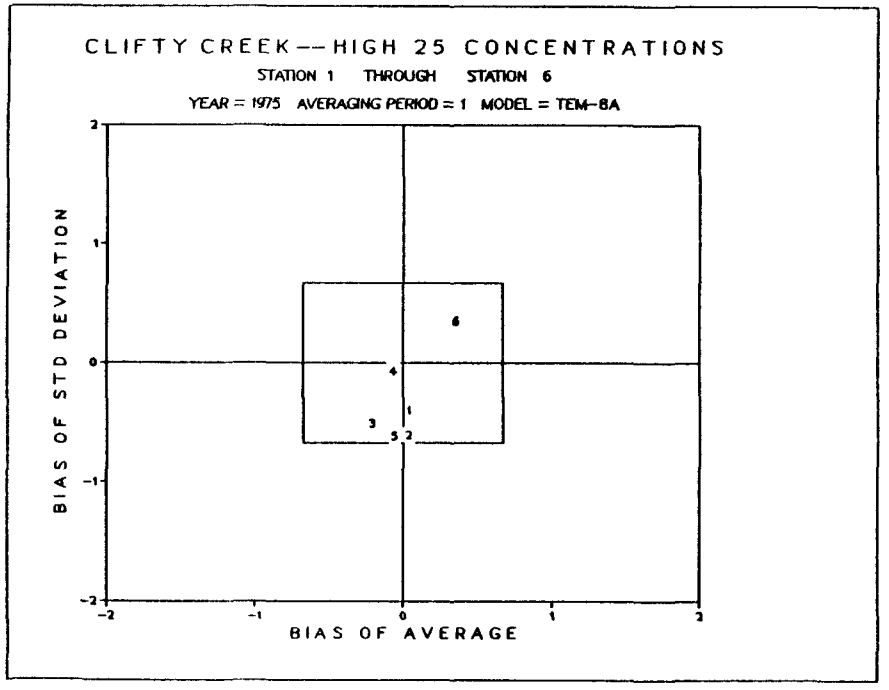


Figure 11. Fractional Bias Plot By Model Using High 25 Values For Each Station

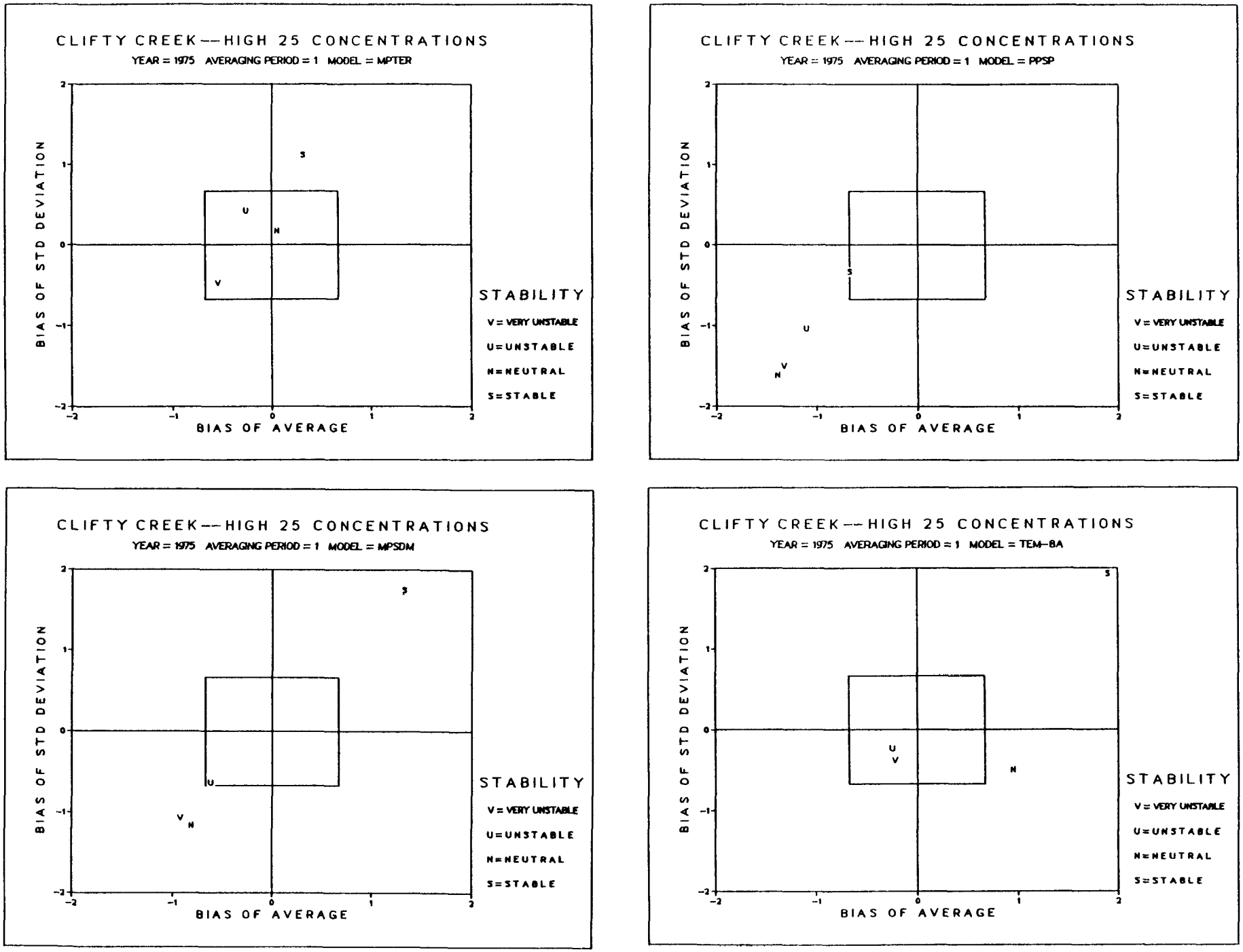
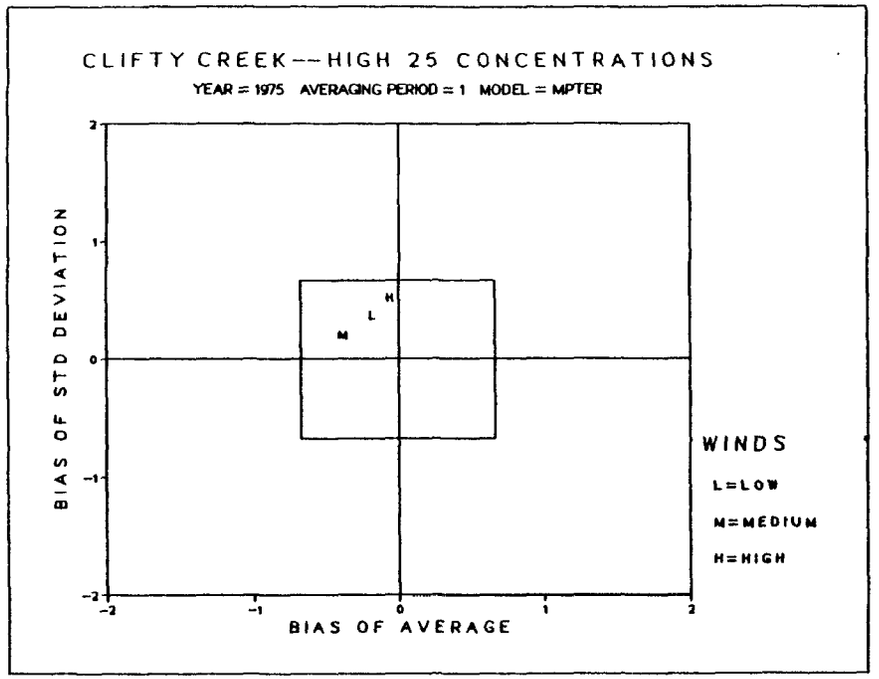
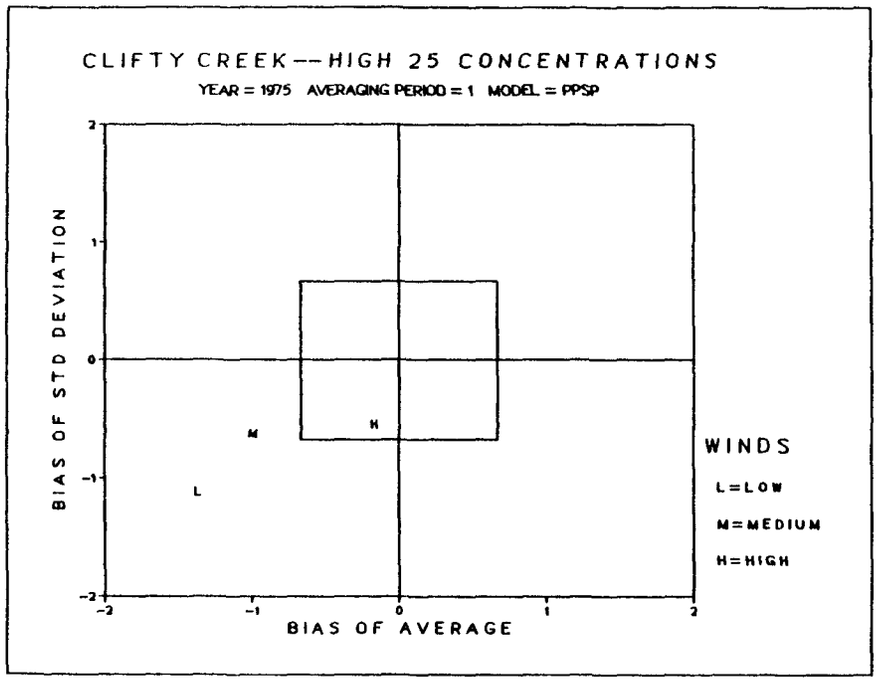
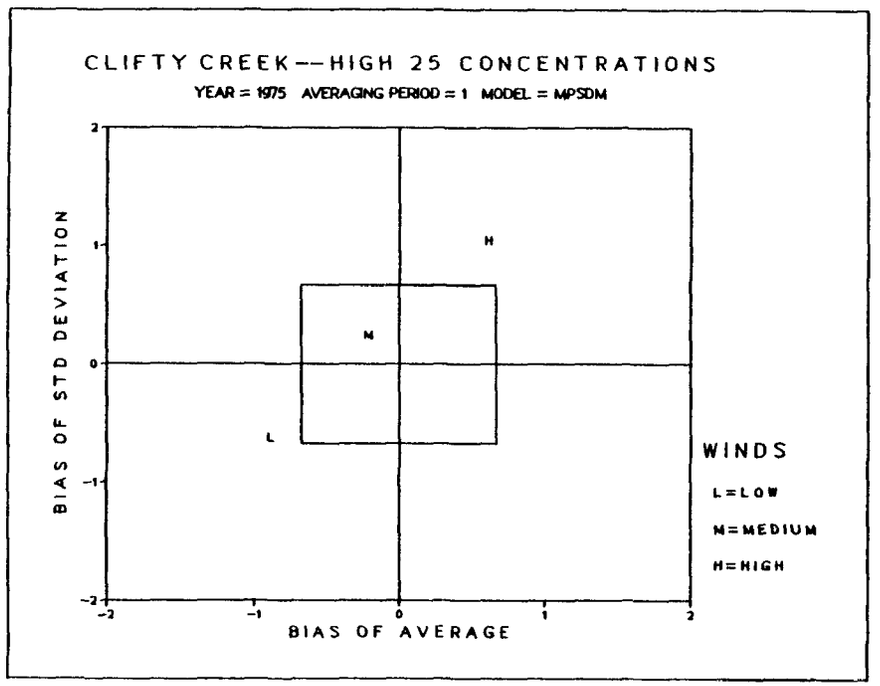


Figure 12. Fractional Bias Plot By Model Using High 25 Values For Each Stability Class







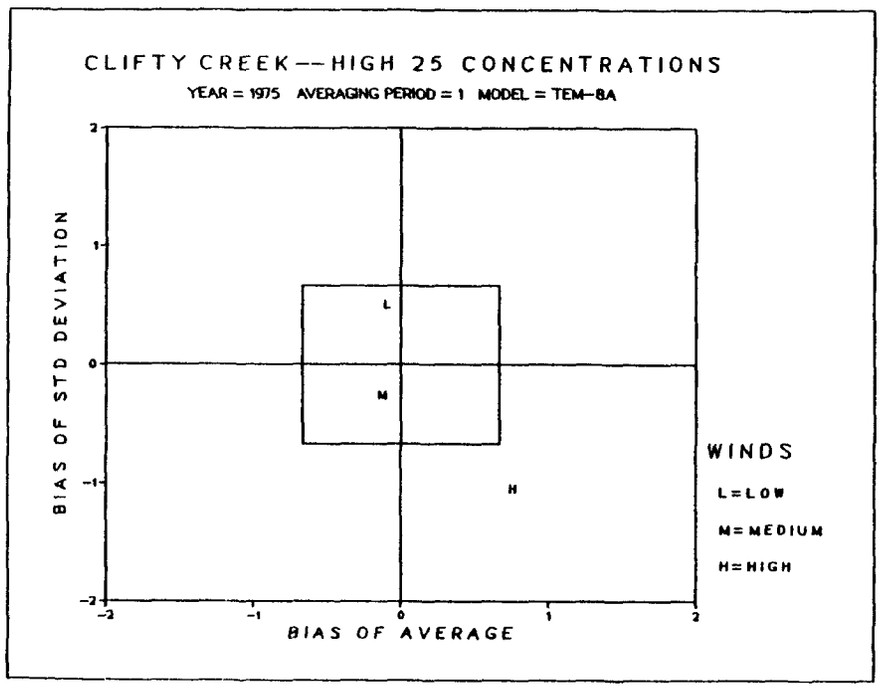
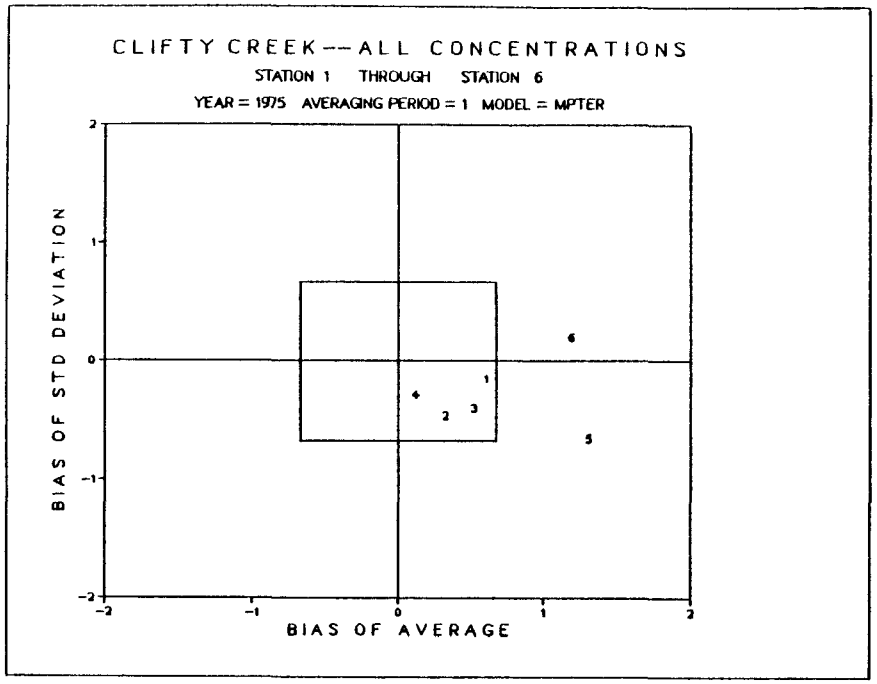
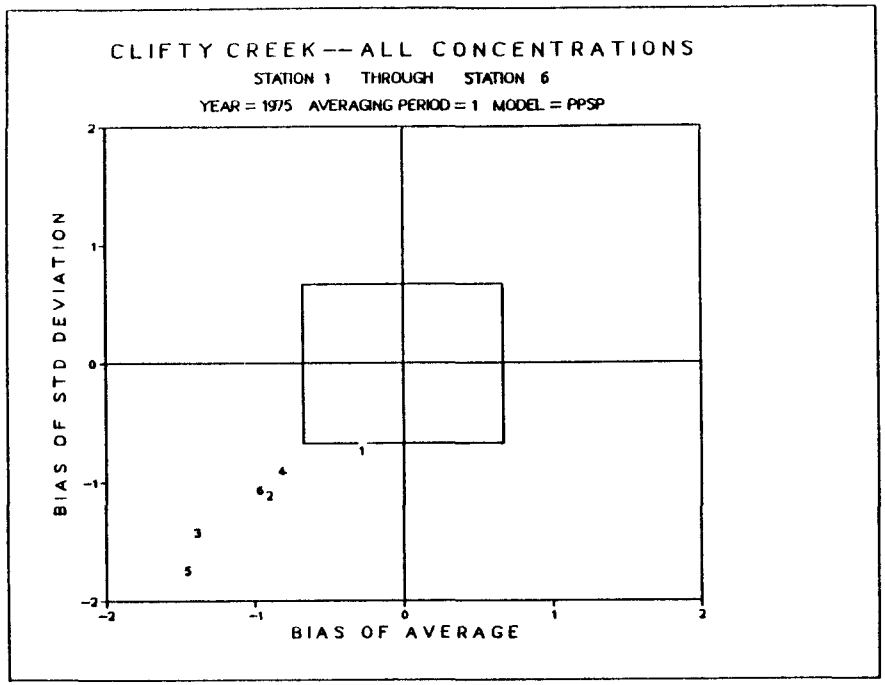
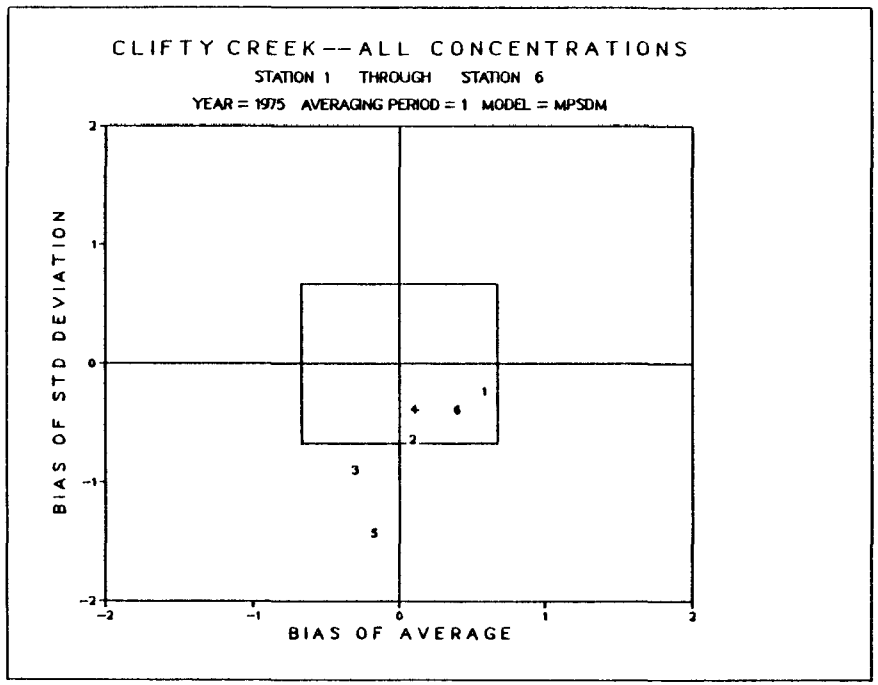


Figure 13. Fractional Bias Plot By Model Using High 25 Values For Each Wind Speed Class







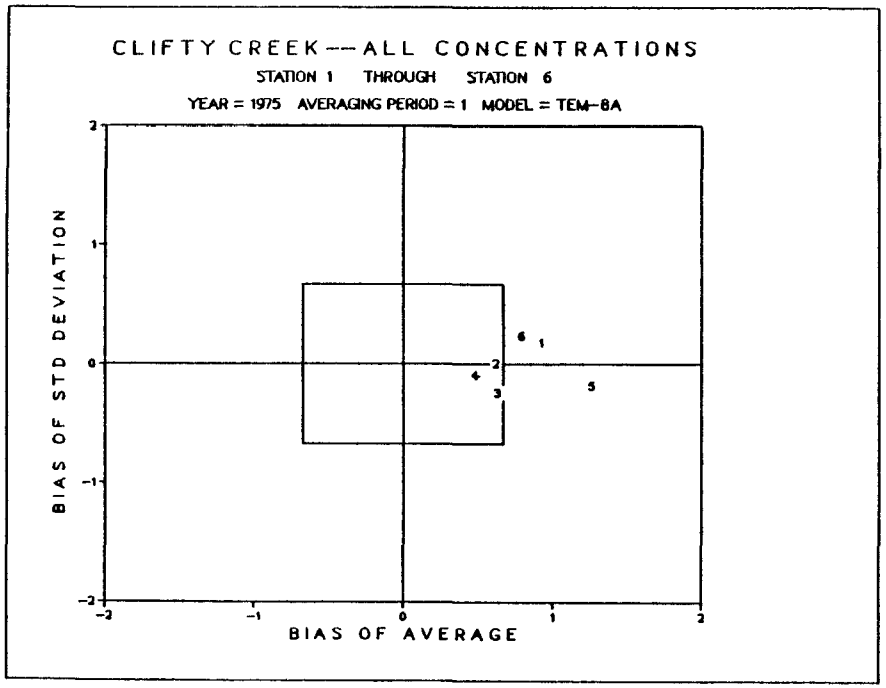
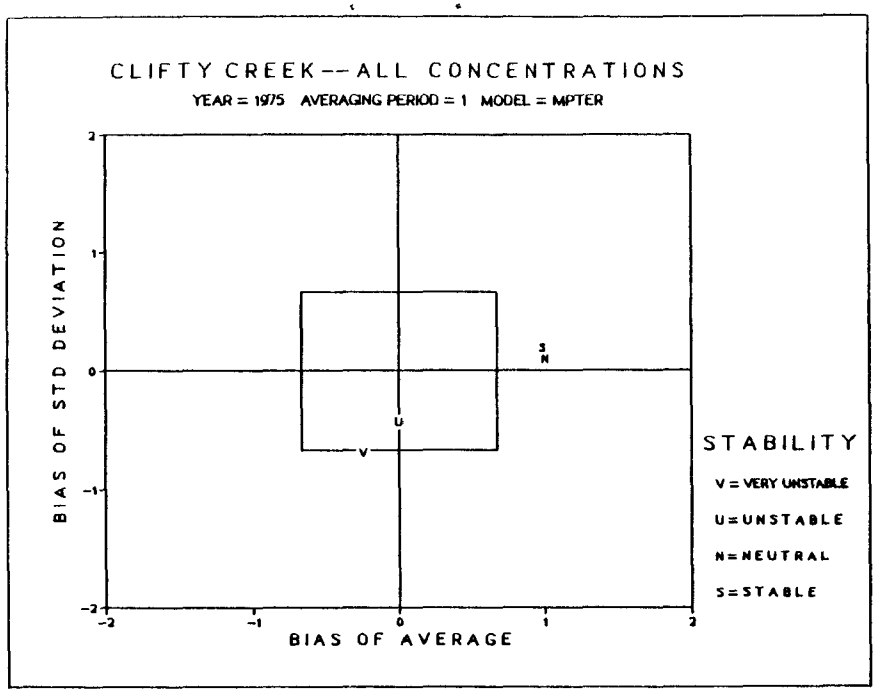
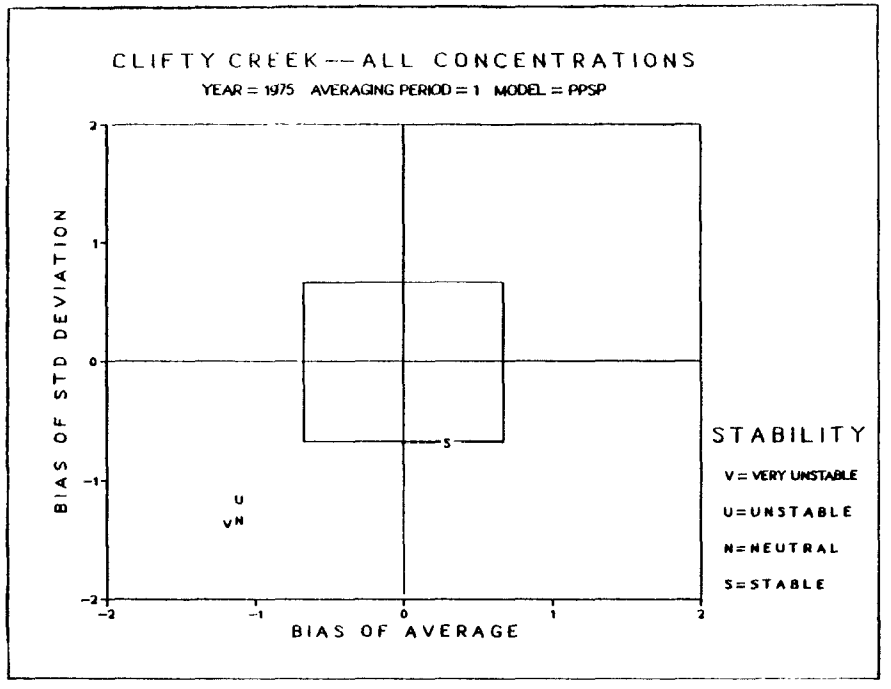
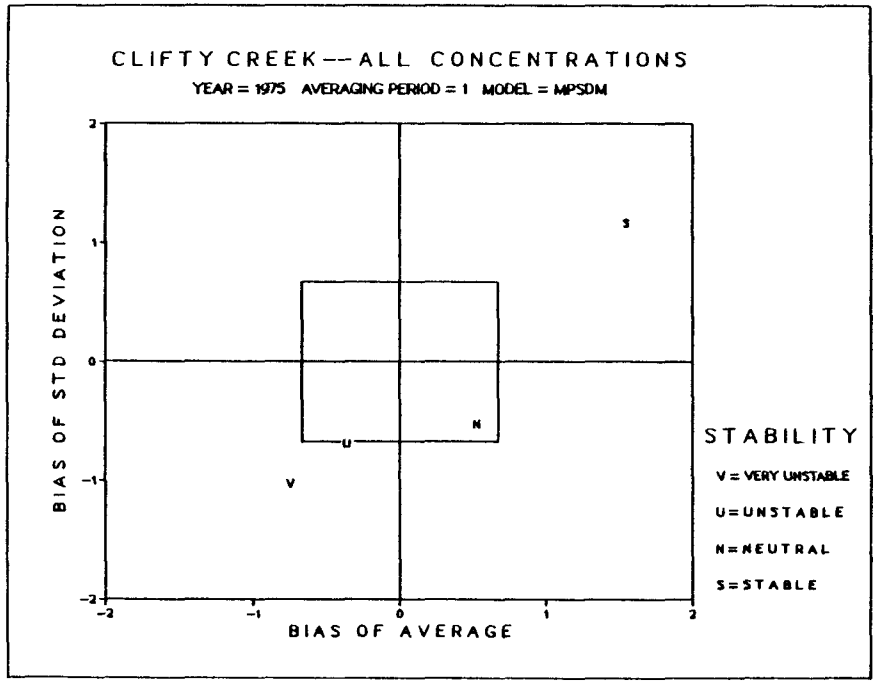


Figure 14. Fractional Bias Plot By Model Using All Paired Values For Each Station







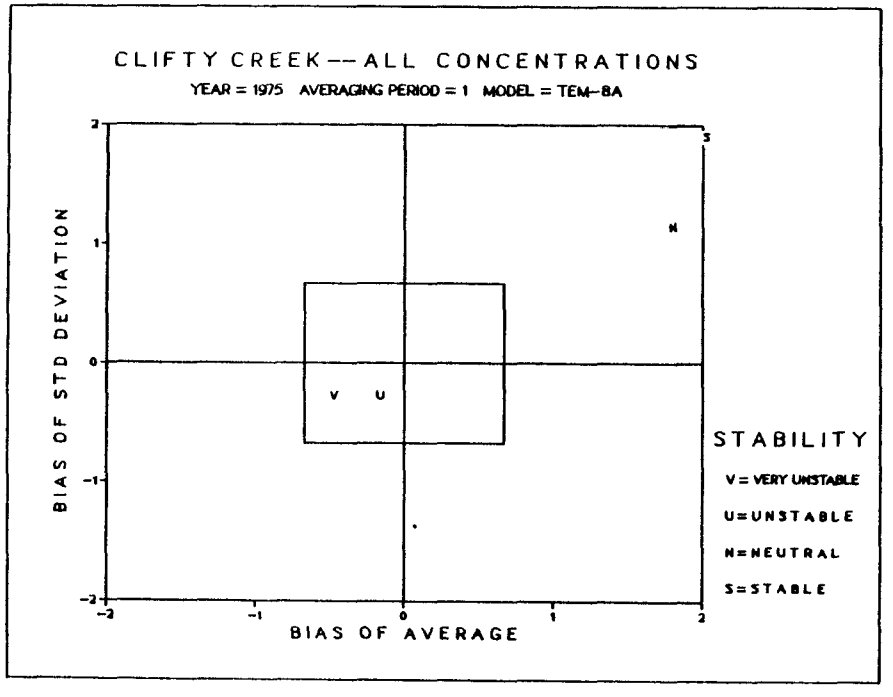
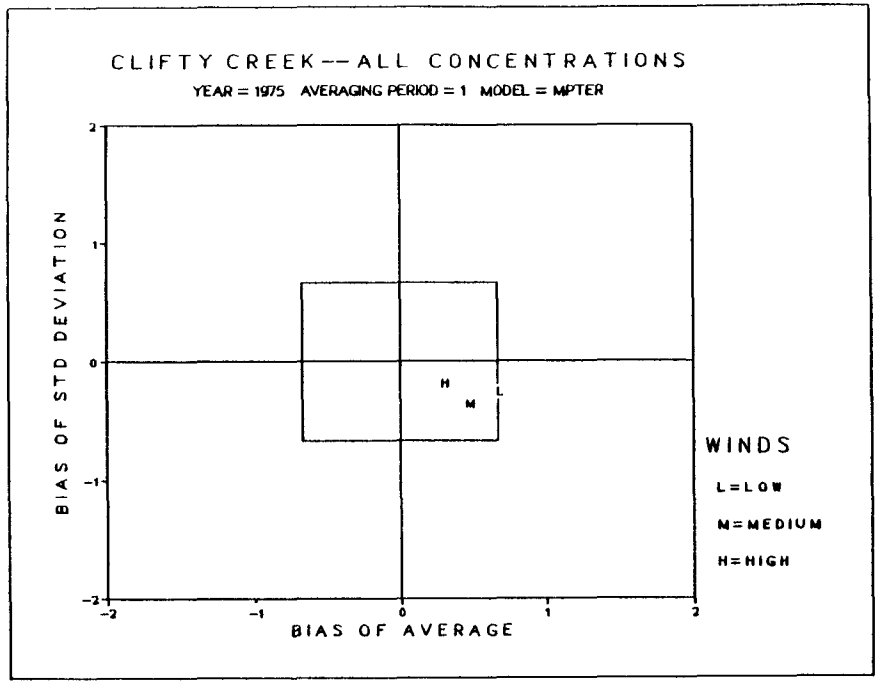
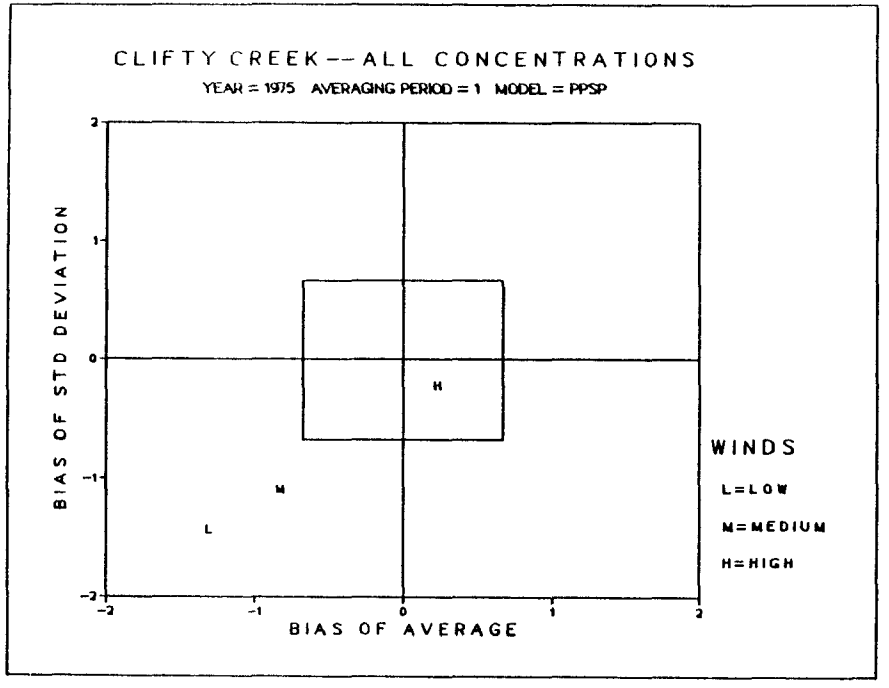
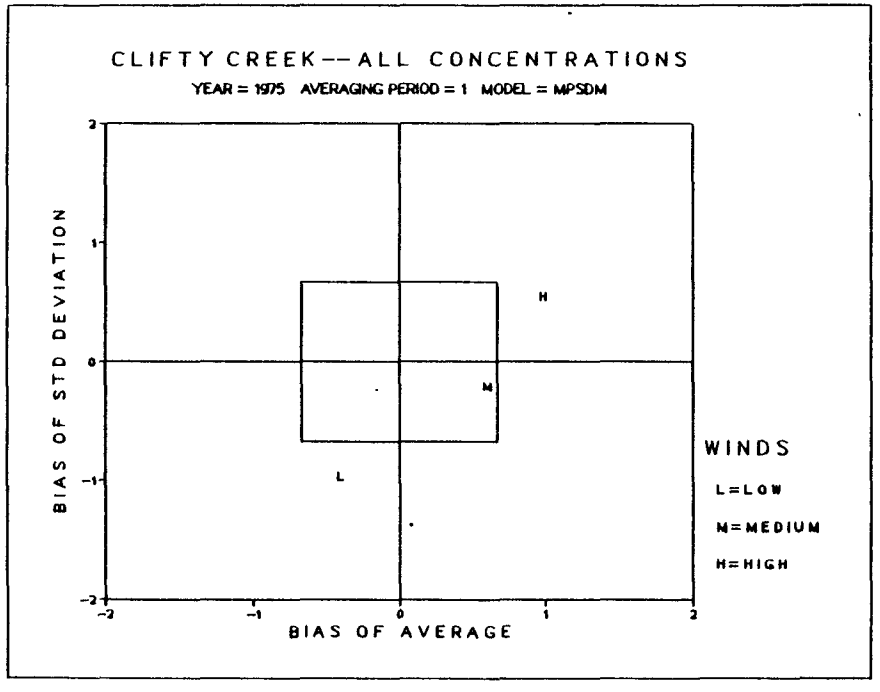


Figure 15. Fractional Bias Plot By Model Using All Paired Values for Each Stability Class







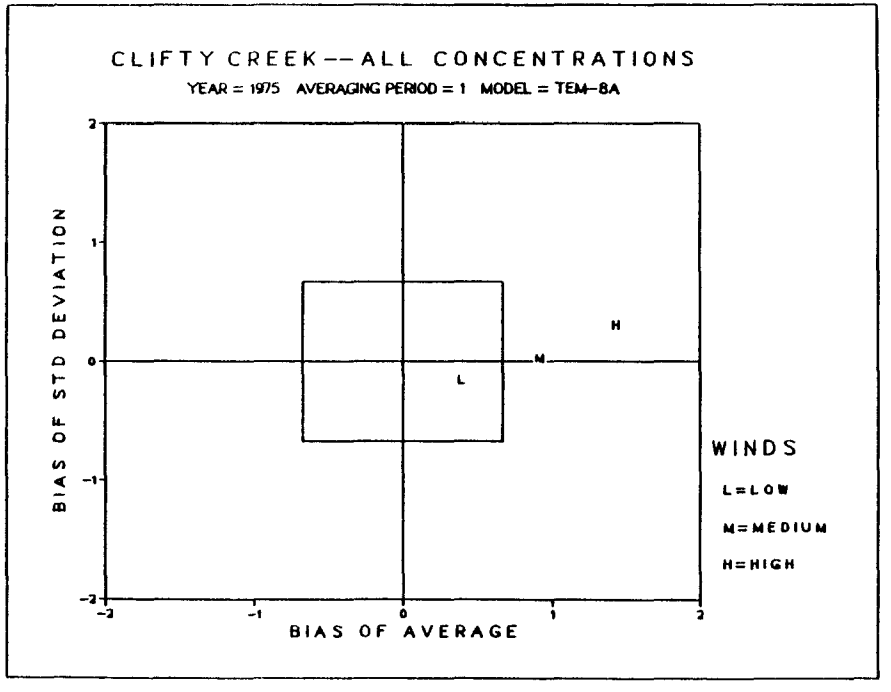


Figure 16. Fractional Bias Plot By Model Using All Paired Values For Each Wind Speed Class

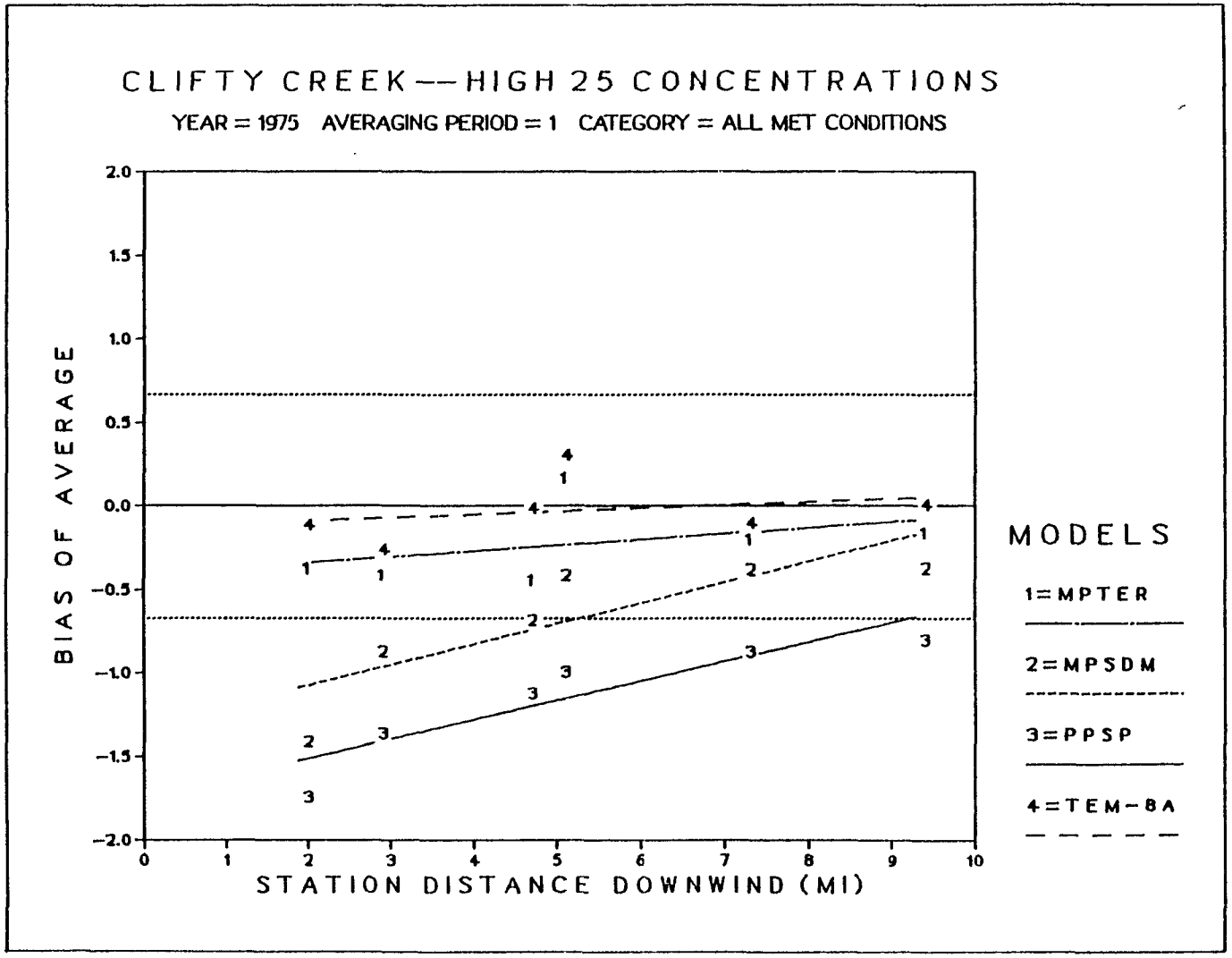
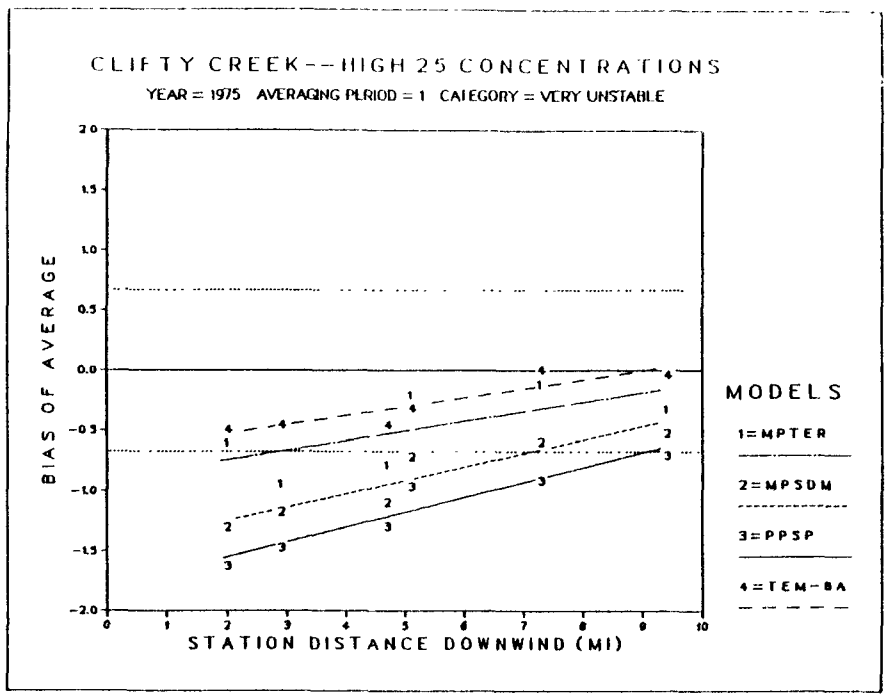
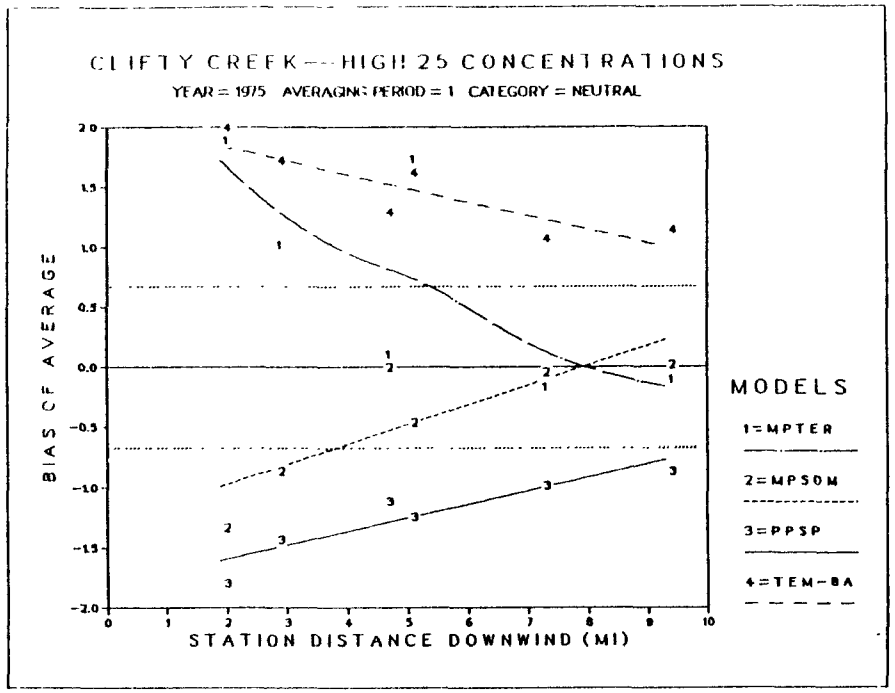
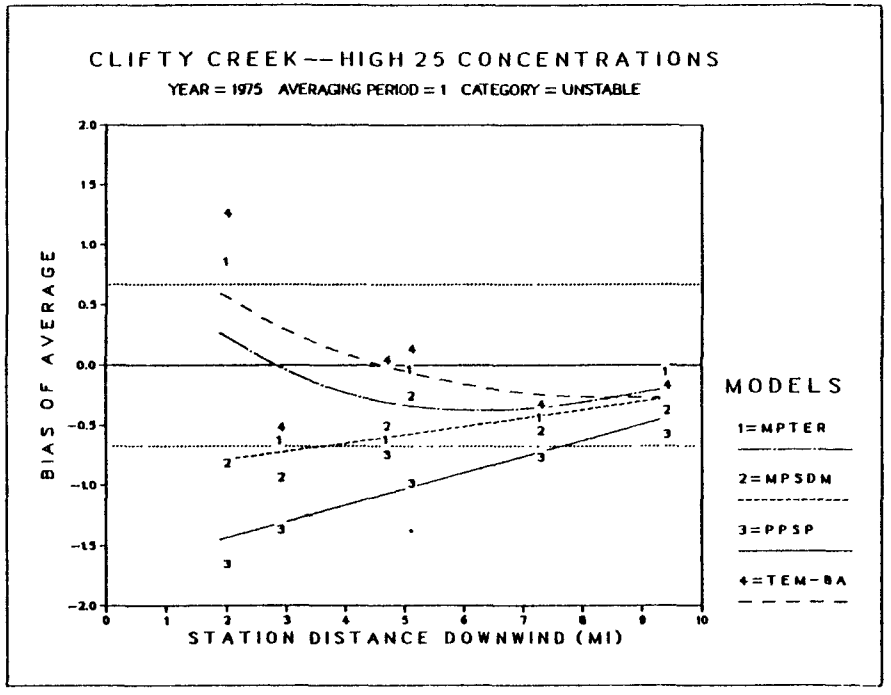


Figure 17. Fractional Bias Of The Average Vs Station Distance Using High 25 Values







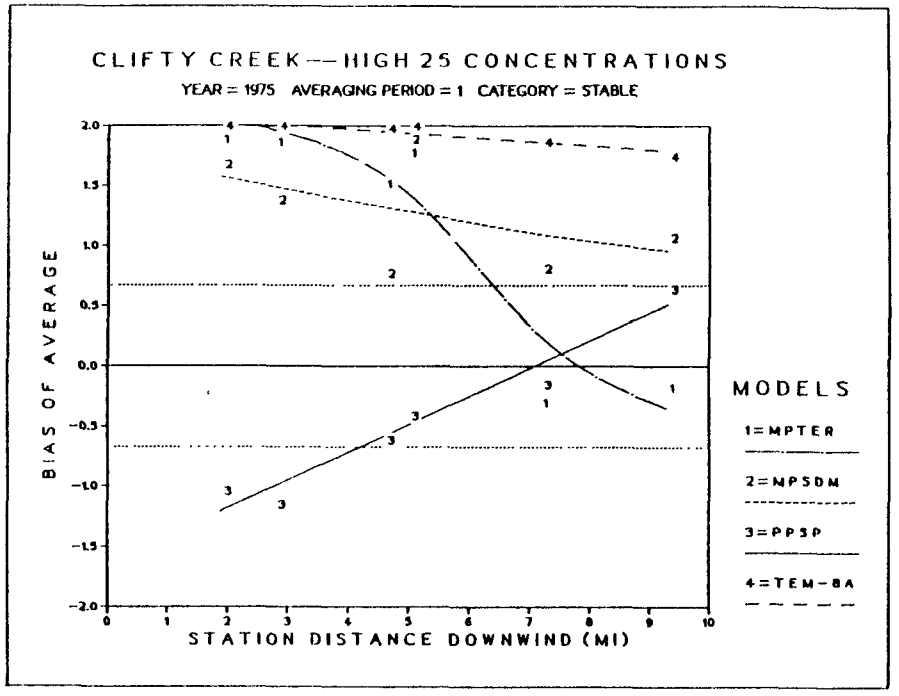
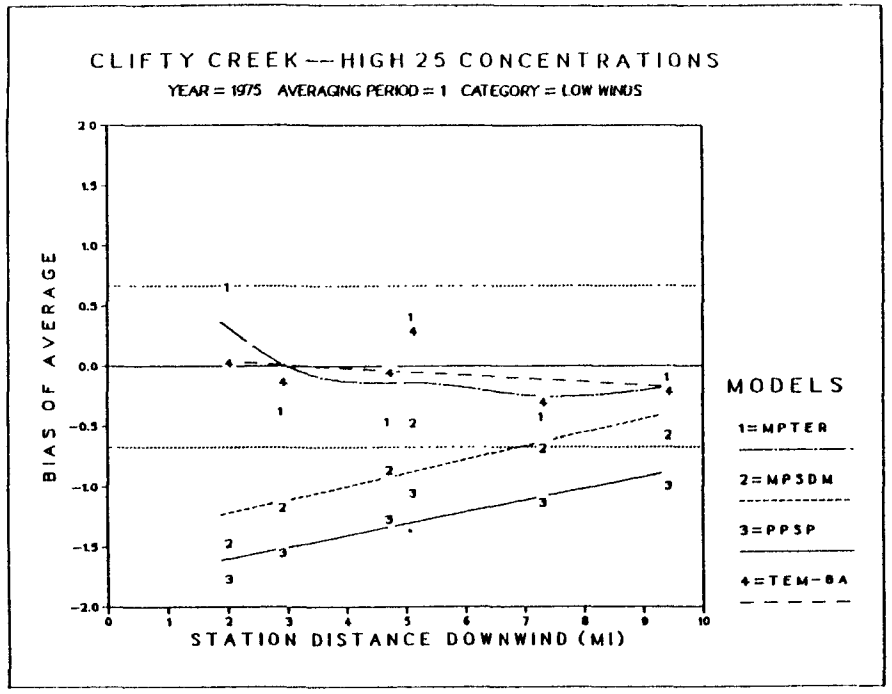
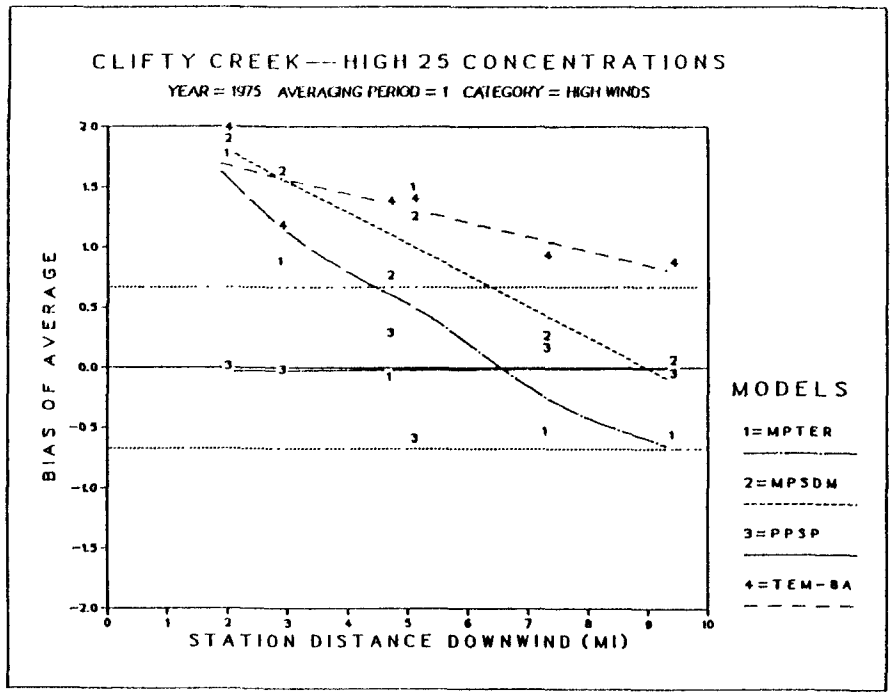


Figure 18. Fractional Bias Of The Average Vs Station Distance By Stability Class Using High 25 Values





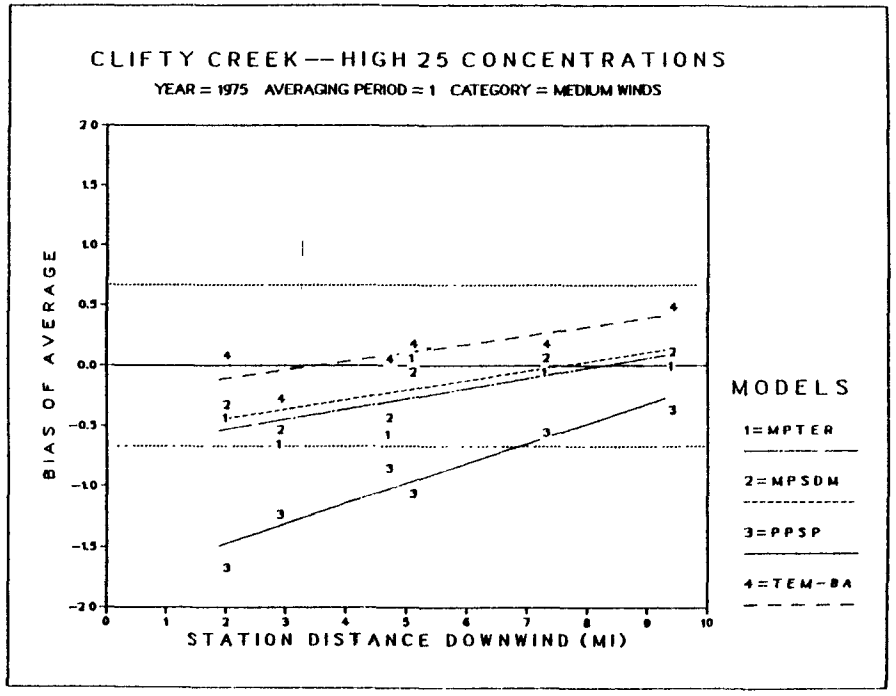


Figure 19. Fractional Bias Of The Average Vs Station Distance By Wind Speed Class Using High 25 Values

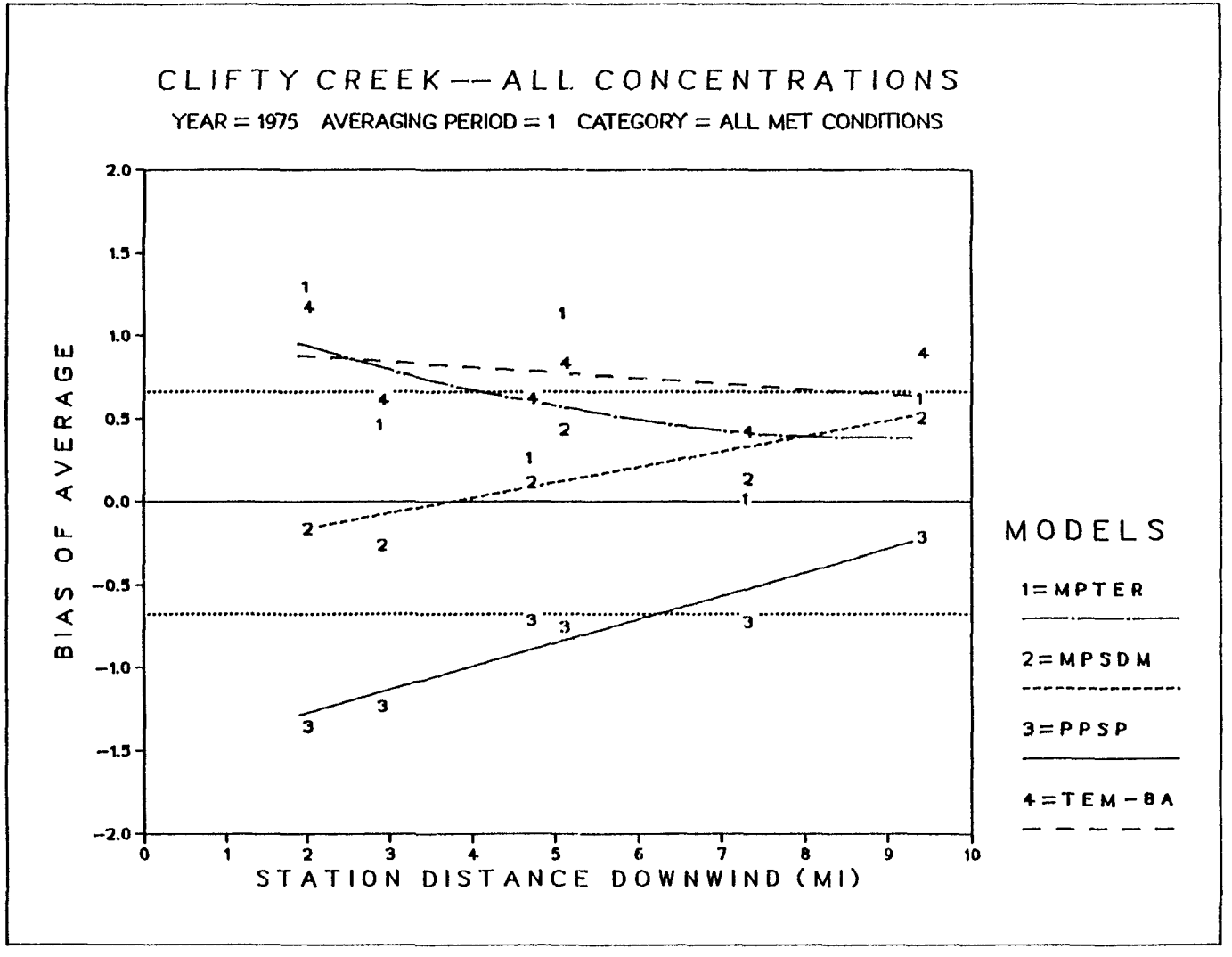
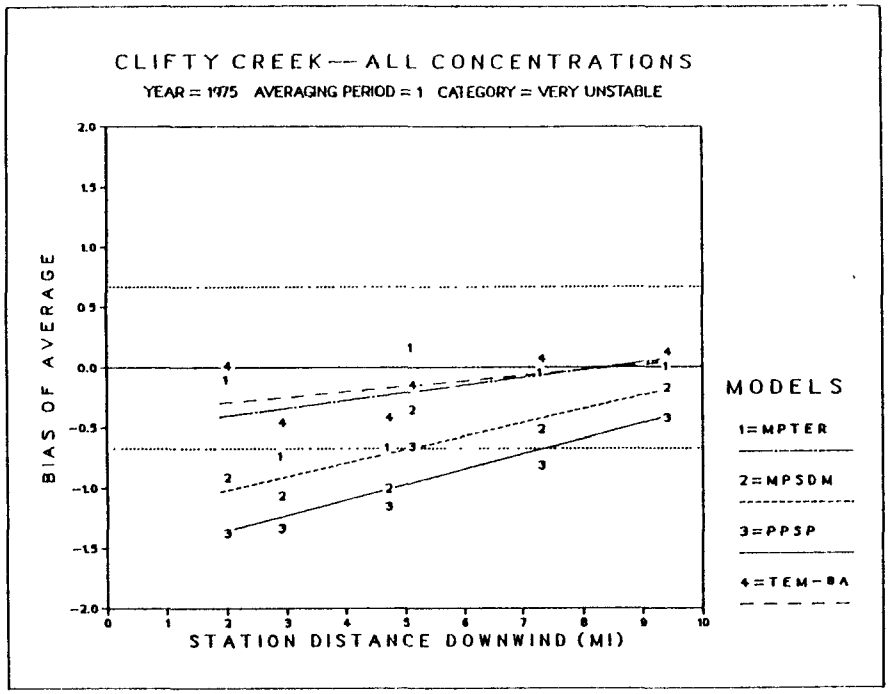
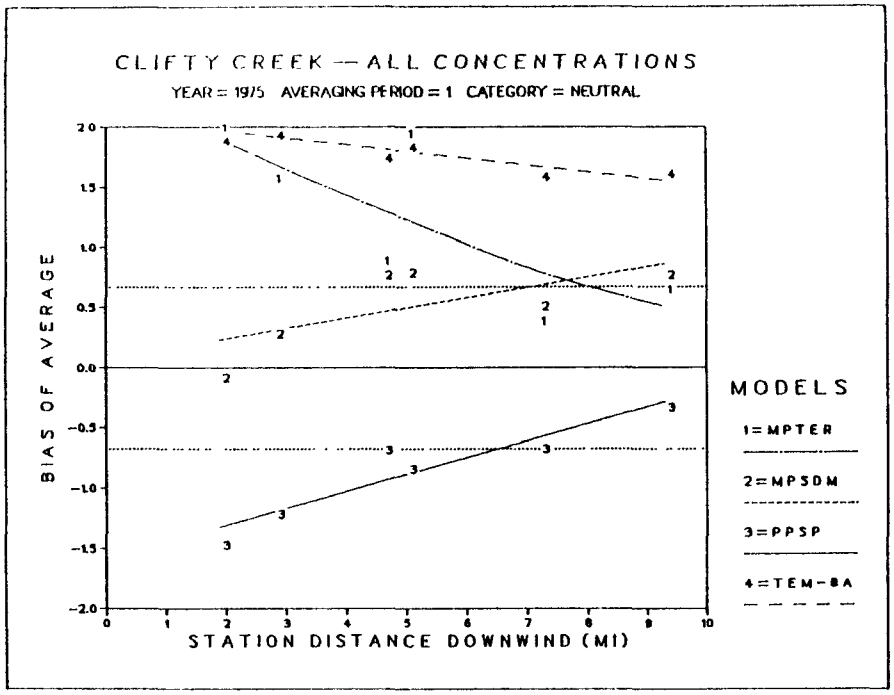
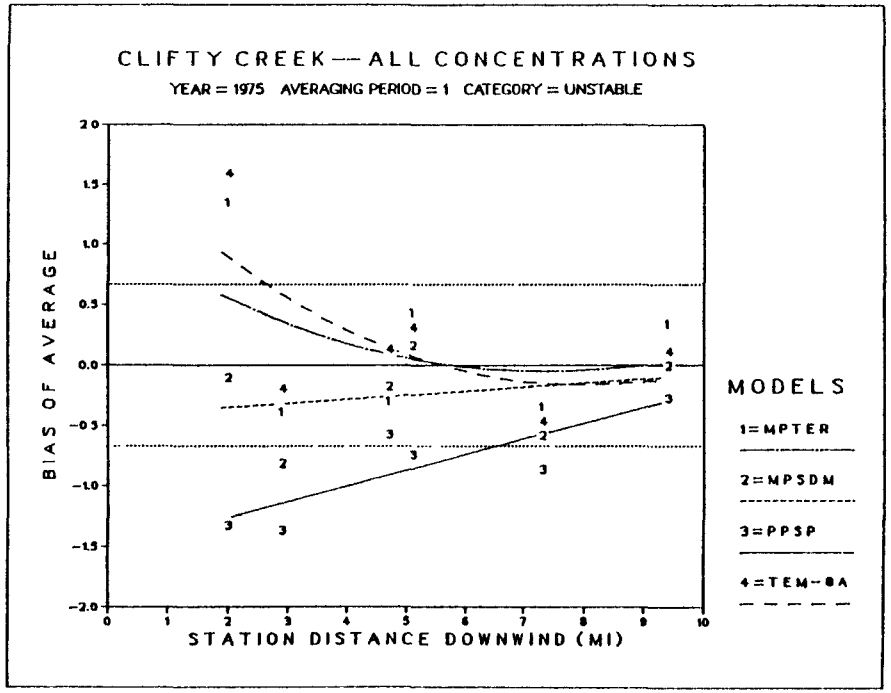


Figure 20. Fractional Bias Of The Average Vs Station Distance Using All Paired Values







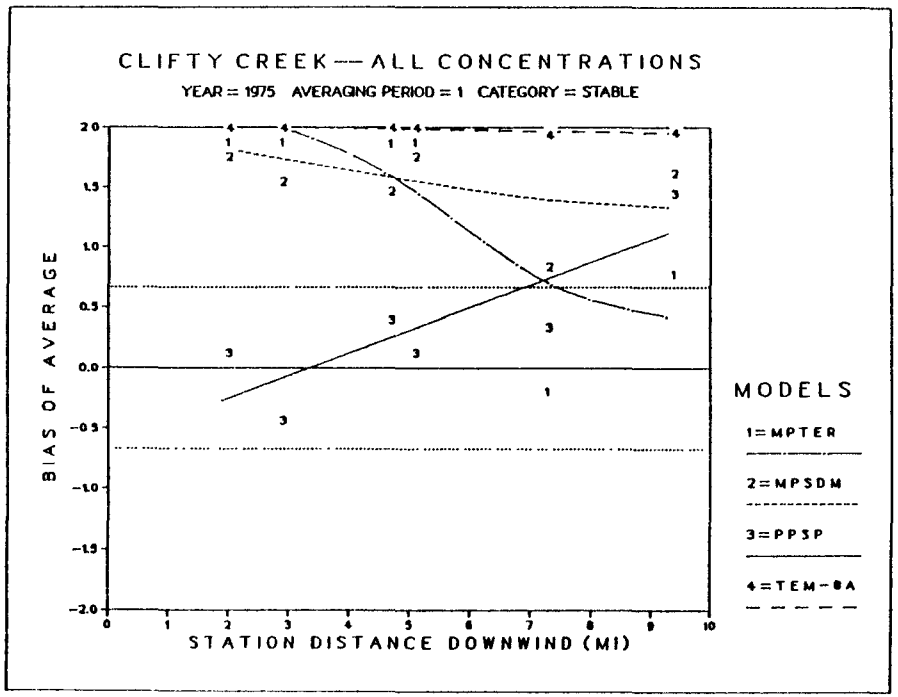
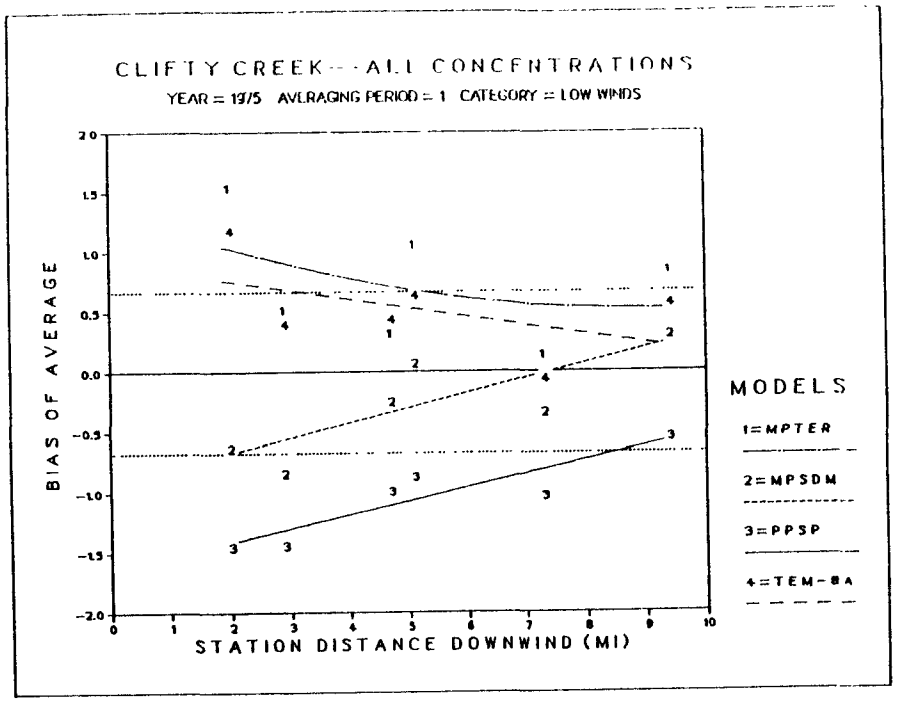
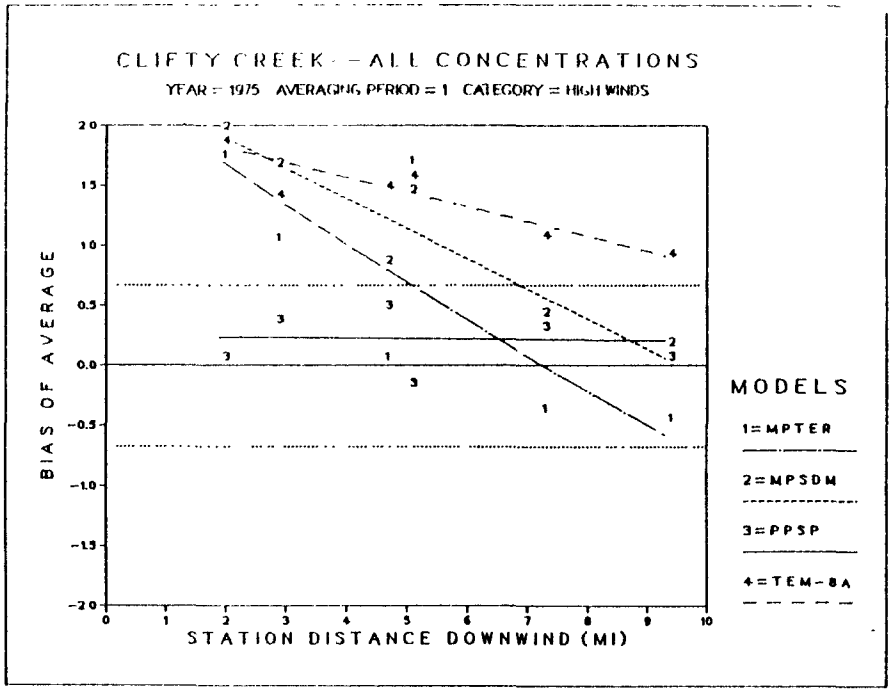


Figure 21. Fractional Bias Of The Average Vs Station Distance By Stability Class Using All Paired Values





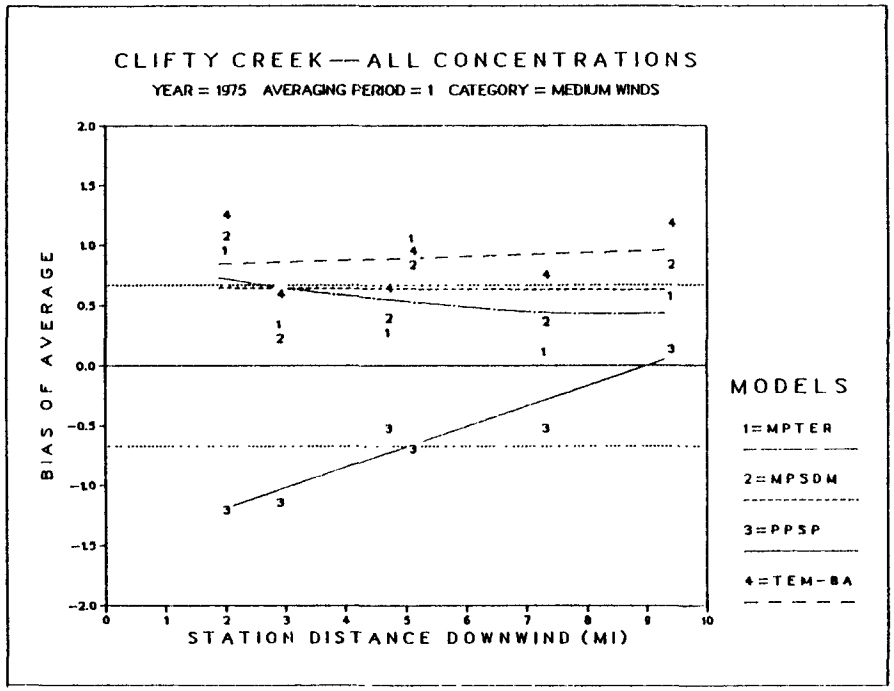


Figure 22. Fractional Bias Of The Average Vs Station Distance By Wind Speed Class Using All Paired Values

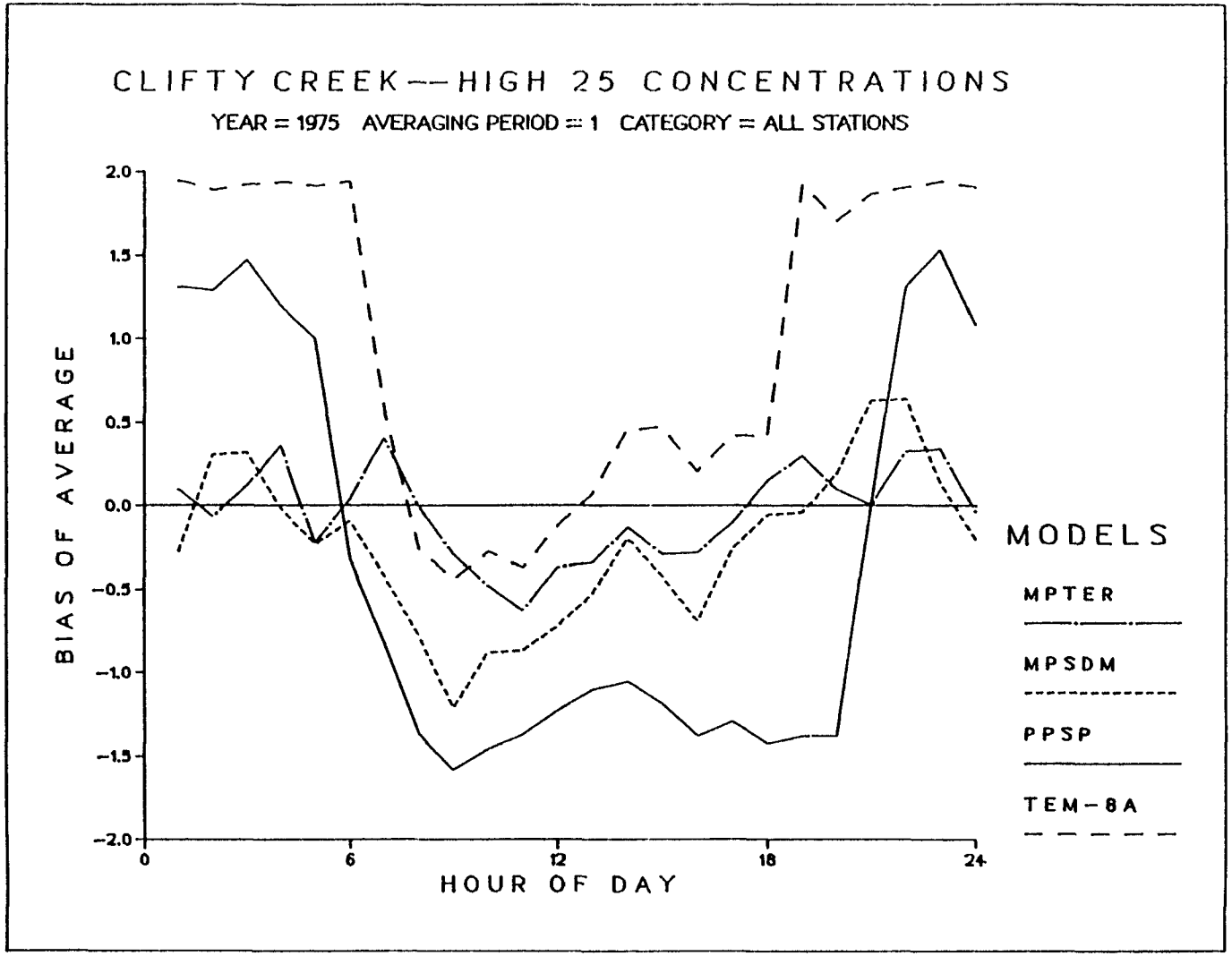
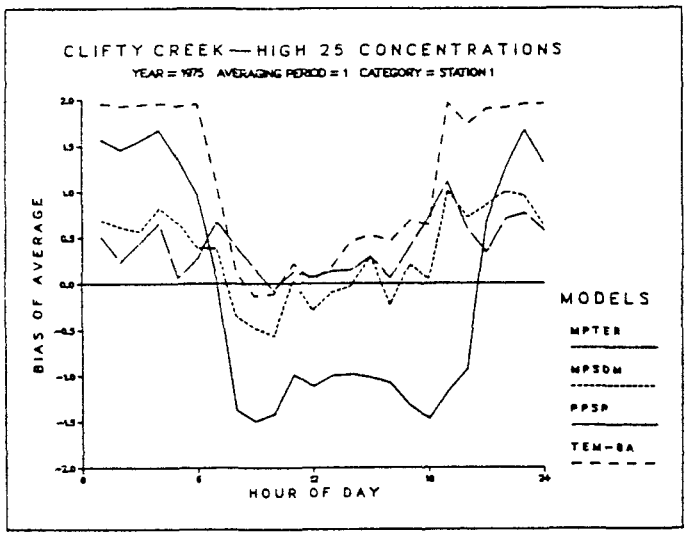
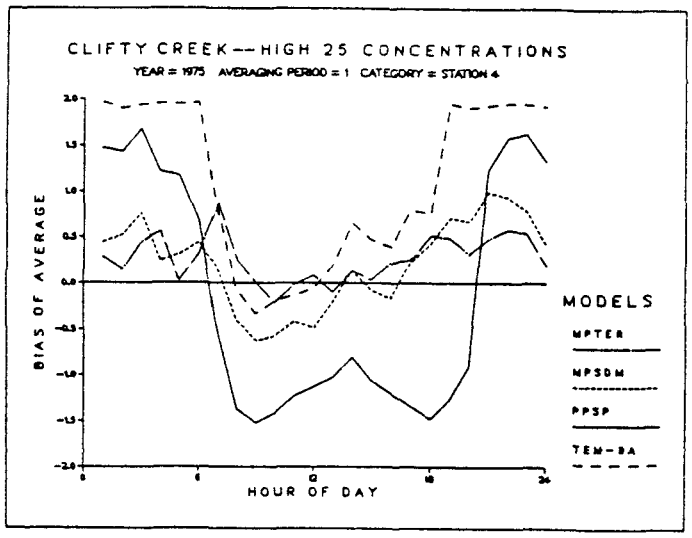
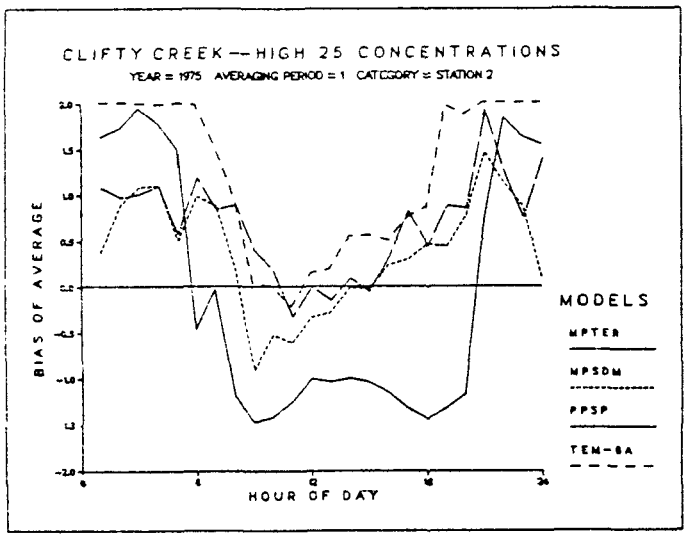
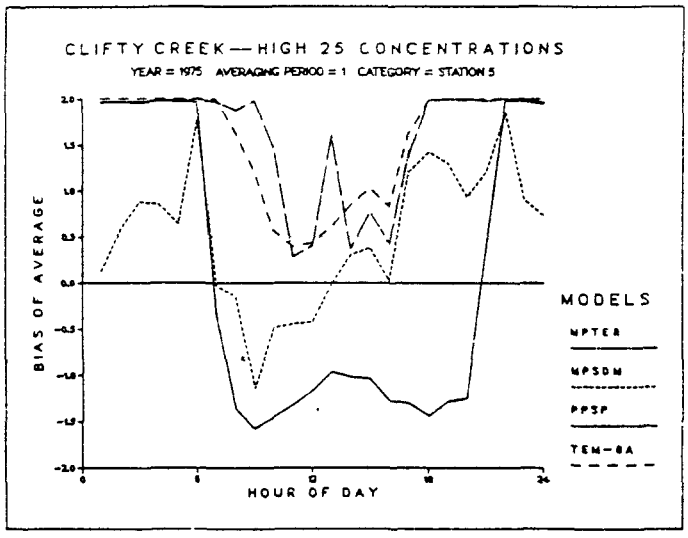


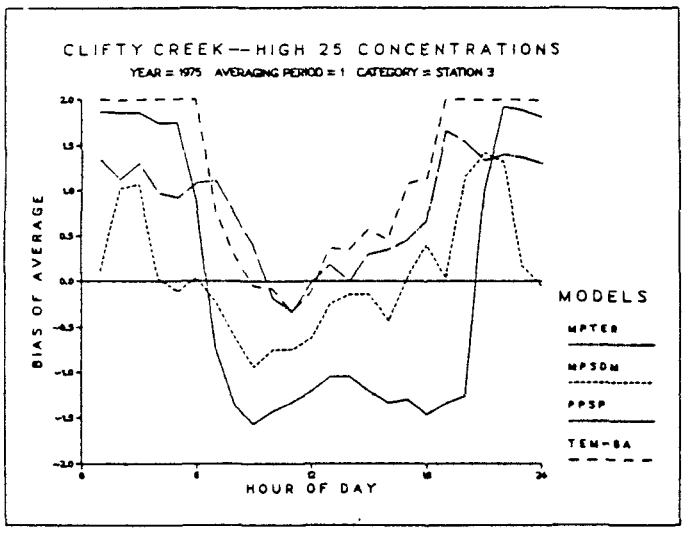
Figure 23. Fractional Bias Of The Average Vs Hour Of The Day Using High 25 Values











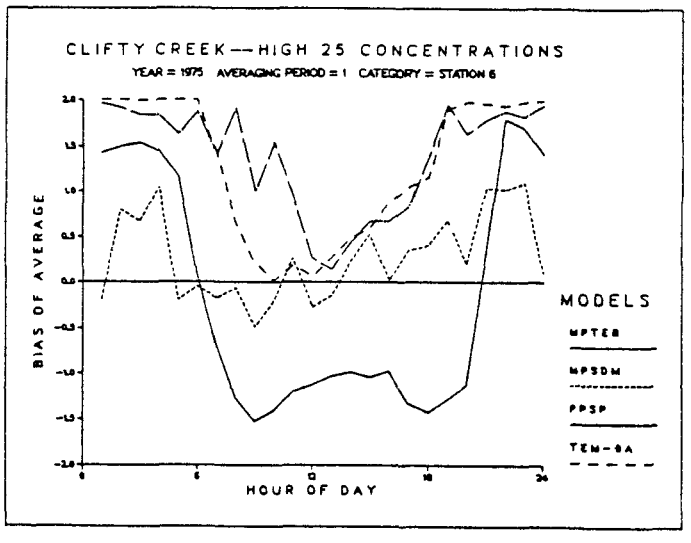


Figure 24. Fractional Bias Of The Average Vs Hour Of The Day By Station Using High 25 Values

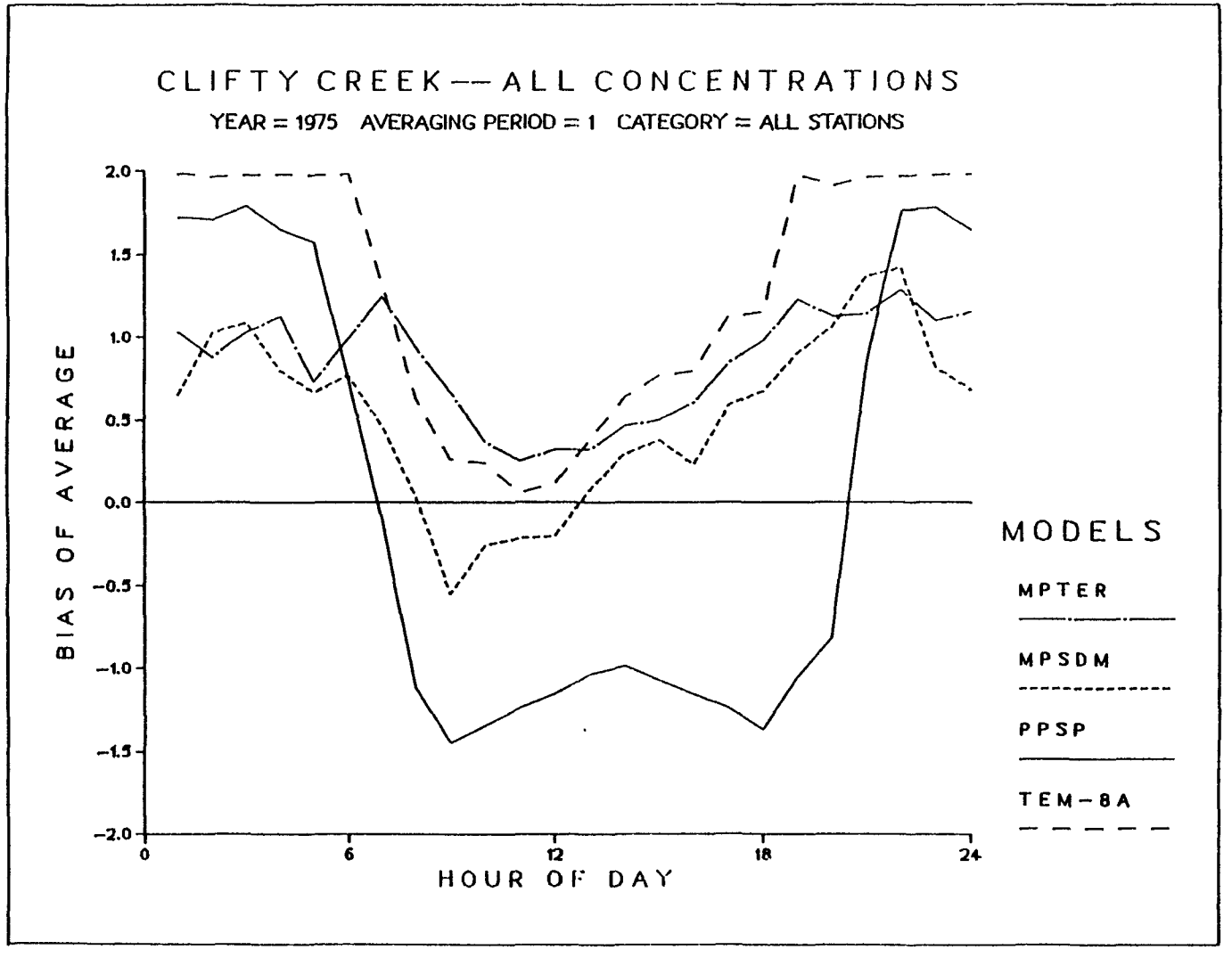
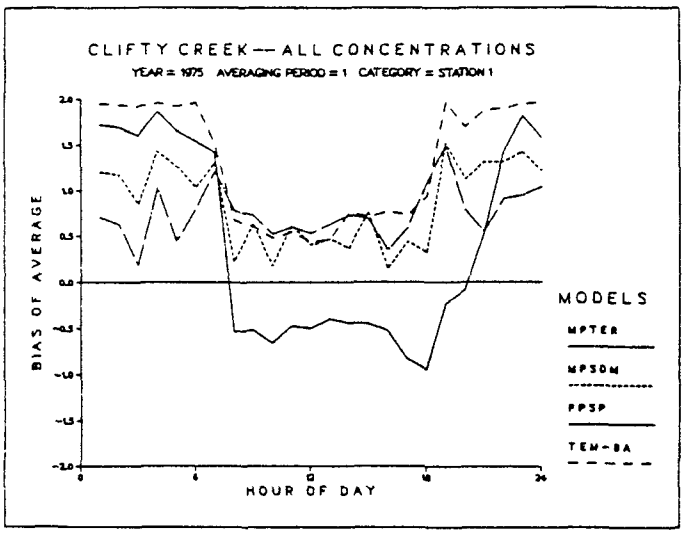
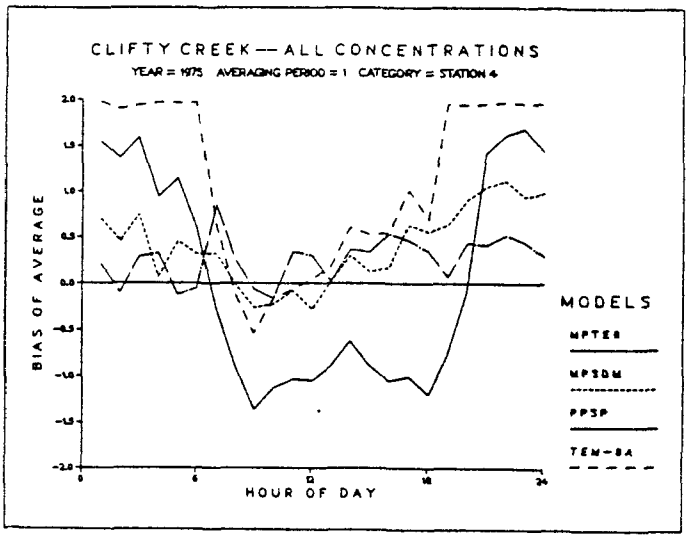
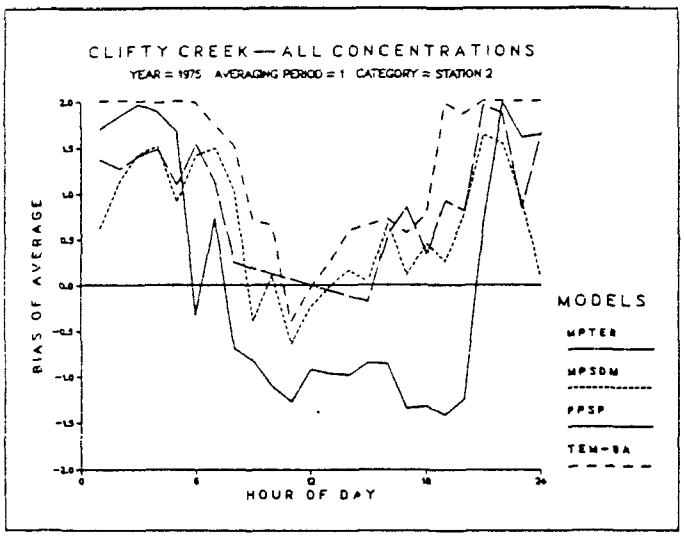
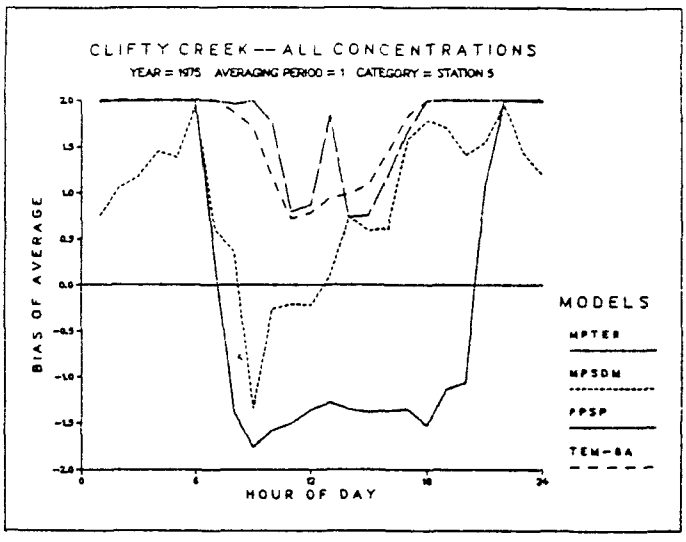


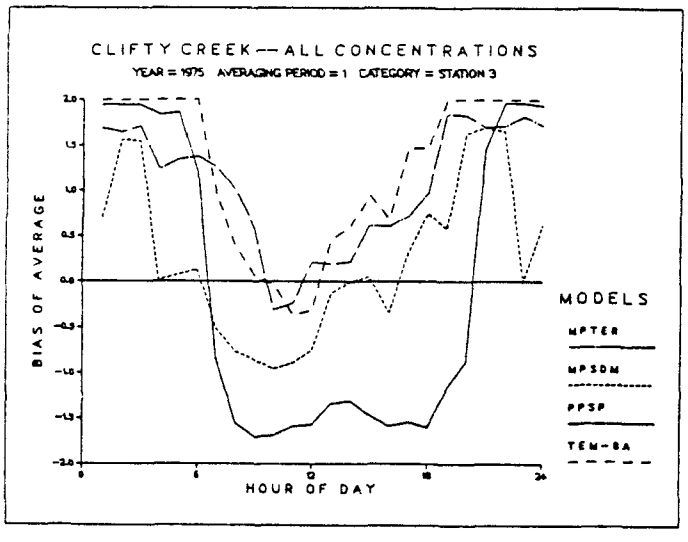
Figure 25. Fractional Bias Of The Average Vs Hour Of The Day Using All Paired Values











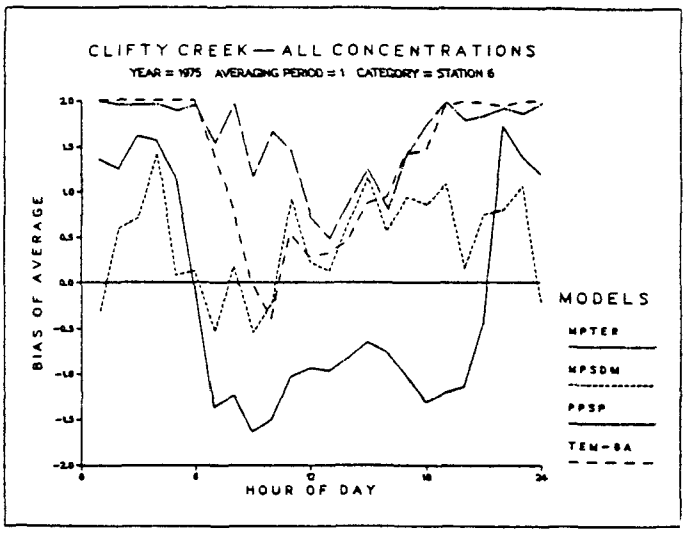


Figure 26. Fractional Bias Of The Average Vs Hour Of The Day by Station Using All Paired Values

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.

4. TITLE AND SUBTITLE
Evaluation of Rural Air Quality Simulation Models
Addendum B: Graphical Display of Model Performance Using the Clifty Creek Data Base

5. REPORT DATE August 1985

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
William M. Cox, Ellen Baldridge
Gerald K. Moss, Joseph A. Tikvart

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Source Receptor Analysis Branch
Monitoring and Data Analysis Division
U.S. Environmental Protection Agency

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS
Same as above

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This addendum uses a variety of graphical formats to display and compare the performance of four rural models using the Clifty Creek data base. The four models included MPTER (EPA), PPSP (Martin Marietta Corp.), MPSDM (ERT) and TEM-8A (Texas Air Control Board). Graphic displays were developed and used for both operational evaluation and diagnostic evaluation purposes. For operational evaluation, simple plots of bias of the standard deviation vs bias of the average proved useful for summarizing and intercomparing the performance of the four rural models. For diagnostic evaluation, selected data subsets by station, meteorological class and hour of the day proved beneficial. Plots of bias of the average vs station downwind distance by stability and wind speed class revealed clear patterns of accentuated underprediction and overprediction for stations closer to the source. PPSP showed a tendancy for decreasing overprediction with increasing station distance for all meteorological subsets while the other three models showed varying patterns depending on the meteorological class. Diurnal plots of the bias of the average vs hour of the day revealed a pattern of underestimation during the nocturnal hours and overestimation during hours of strong solar radiation with MPSDM and MPTER showing the least overall bias throughout the day.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Mathematical Modeling Meteorology Sulfur Dioxide Statistical Measure Performance Evaluation Graphic Display	Air Quality Impact Assessment	

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

INSTRUCTIONS

1. **REPORT NUMBER**
   Insert the EPA report number as it appears on the cover of the publication.

2. **LEAVE BLANK**

3. **RECIPIENTS ACCESSION NUMBER**
   Reserved for use by each report recipient.

4. **TITLE AND SUBTITLE**
   Title should indicate clearly and briefly the subject coverage of the report, and be displayed prominently. Set subtitle, if used, in smaller type or otherwise subordinate it to main title. When a report is prepared in more than one volume, repeat the primary title, add volume number and include subtitle for the specific title.

5. **REPORT DATE**
   Each report shall carry a date indicating at least month and year. Indicate the basis on which it was selected *(e.g., date of issue, date of approval, date of preparation, etc.)*.

6. **PERFORMING ORGANIZATION CODE**
   Leave blank.

7. **AUTHOR(S)**
   Give name(s) in conventional order *(John R. Doe, J. Robert Doe, etc.)*. List author's affiliation if it differs from the performing organization.

8. **PERFORMING ORGANIZATION REPORT NUMBER**
   Insert if performing organization wishes to assign this number.

9. **PERFORMING ORGANIZATION NAME AND ADDRESS**
   Give name, street, city, state, and ZIP code. List no more than two levels of an organizational hirearchy.

10. **PROGRAM ELEMENT NUMBER**
    Use the program element number under which the report was prepared. Subordinate numbers may be included in parentheses.

11. **CONTRACT/GRANT NUMBER**
    Insert contract or grant number under which report was prepared.

12. **SPONSORING AGENCY NAME AND ADDRESS**
    Include ZIP code.

13. **TYPE OF REPORT AND PERIOD COVERED**
    Indicate interim final, etc., and if applicable, dates covered.

14. **SPONSORING AGENCY CODE**
    Insert appropriate code.

15. **SUPPLEMENTARY NOTES**
    Enter information not included elsewhere but useful, such as: Prepared in cooperation with, Translation of, Presented at conference of, To be published in, Supersedes, Supplements, etc.

16. **ABSTRACT**
    Include a brief *(200 words or less)* factual summary of the most significant information contained in the report. If the report contains a significant bibliography or literature survey, mention it here.

17. **KEY WORDS AND DOCUMENT ANALYSIS**
    (a) DESCRIPTORS - Select from the Thesaurus of Engineering and Scientific Terms the proper authorized terms that identify the major concept of the research and are sufficiently specific and precise to be used as index entries for cataloging.

    (b) IDENTIFIERS AND OPEN-ENDED TERMS - Use identifiers for project names, code names, equipment designators, etc. Use open-ended terms written in descriptor form for those subjects for which no descriptor exists.

    (c) COSATI FIELD GROUP - Field and group assignments are to be taken from the 1965 COSATI Subject Category List. Since the majority of documents are multidisciplinary in nature, the Primary Field/Group assignment(s) will be specific discipline, area of human endeavor, or type of physical object. The application(s) will be cross-referenced with secondary Field/Group assignments that will follow the primary posting(s).

18. **DISTRIBUTION STATEMENT**
    Denote releasability to the public or limitation for reasons other than security for example "Release Unlimited." Cite any availability to the public, with address and price.

19. & 20. **SECURITY CLASSIFICATION**
    DO NOT submit classified reports to the National Technical Information service.

21. **NUMBER OF PAGES**
    Insert the total number of pages, including this one and unnumbered pages, but exclude distribution list, if any.

22. **PRICE**
    Insert the price set by the National Technical Information Service or the Government Printing Office, if known.