DOE/ER/60557--1

DE91 007575

EXPLORATORY DATA ANALYSIS ON DATA GENERATED IN THE DOE SUBSURFACE MICROBIOLOGY PROGRAM

(CONTRACT # DE FG02-87ER60557)

FINAL REPORT
JUNE 1990

DISCLAIMER

Principal Investigator:
ROBERT R. MEGLEN

Center for Environmental Sciences
University of Colorado at Denver
Campus Box 136
1200 Larimer St.
Denver, CO 80204

CONTENTS

FINAL REPORT
SUMMARY

EXPLORATORY DATA ANALYSIS ON DATA GENERATED IN THE DOE SUBSURFACE MICROBIOLOGY PROGRAM

(CONTRACT # DE FG02-87ER60557)

ROBERT R. MEGLEN

Center for Environmental Sciences
University of Colorado at Denver
Campus Box 136
1200 Larimer St.
Denver, CO 80204

INTRODUCTION

The preliminary results from the innovative subsurface microbiology research program indicate that new data on the nature of the link between the geosphere and biosphere have been generated. The diversity of scientific disciplines represented in the subsurface microbiology program reflects the complexity of the system under study. The research carried out by national laboratory and university research scientists is addressing fundamental questions about the abundance of microorganisms and factors controlling microbial activity in the complex subsurface hydrologic and geochemical environment. Long-term implications of this research for mitigating contamination are clear and researchers share the broader objective of linking the basic science with applied work.

RATIONALE FOR THIS WORK AND STATEMENT OF PROBLEM

It was apparent from the original subsurface microbiology experimental designs and sampling programs that key elements of the collaborative effort were in place. However, Dr. Frank Wobber of the DOE Ecological Research Division, Office of Health and Environmental Research suggested that the complexity of the system under study and the magnitude of the data analysis task required additional attention be given to a formal data analysis plan. He suggested an innovative multivariate approach be taken and that a plan capable of integrating data on hydrological, mineralogical, microbiological and chemical processes be undertaken. Based upon his initial conception, this work was undertaken in an attempt to assist in the exploratory data analysis and to provide investigator's with a mechanism by which data could be summarized for their use.

Data on a large number of variables have been acquired. However, massive quantities of data and collaboration alone will not will not ensure that an adequate mechanistic description of the system will obtain. The problem is that in complex environmental systems significant patterns in the data are not always obvious when one examines the data by conventional data analysis methods, one variable at a time. For data interpretation purposes, complexity may be viewed as interactions (correlations or covariances) among many variables. Interactions among the measured variables tend to dominate the data in complex systems and this useful information is not extracted by univariate approaches. The existence of many variables from each of several measurement domains (microbiology, chemistry, geology, hydrology, etc.) suggested that multivariate statistical techniques should be used to examine the subsurface microbiology data. We emphasize that data are not information. Powerful interpretive aids that match the sophistication of the experimental

design and measurement instruments assist in the task of converting data into information and finally into knowledge of the system.

The strong sampling design and potential high quality of the existing data provided an extremely valuable data base for exploratory data analysis. Several features of the subsurface environment mandate multivariate examination. Vertical and lateral transport of carbon and other potential microbial nutrients will clearly depend upon hydrologic and physical properties of texture/porosity. In turn, chemistry of pore waters will depend upon mineralogic characteristics, surface activities, and local hydrologic characteristics. Insight regarding the system clearly requires the ability to obtain correlations of many covariates between geological strata. In addition, potential seasonal variability and other stochastic sources of variance often confound identification of significant system interactions.

The work performed under this contract was designed to provide investigators with an exploratory data analysis of the subsurface microbiology data. Principal component analysis and allied multivariate techniques of computer assisted pattern recognition were used to identify key features and to summarize the multivariate character of the data.

The objectives of the work were:

to provide a multivariate summary of the data base for collaborators;

to provide an interpretive communications link for collaborators;

to identify key variables that characterize the subsurface;

to identify significant microbiological/geological/hydrological/chemical groupings of the subsurface samples.

THE MODE OF OPERATION - GENERAL DESCRIPTION

An effort to obtain cooperation from individual investigators was undertaken in order to create a central data base for use in the multivariate exploratory data analysis. In return for their cooperation investigators were provided with multivariate summaries of their data. In addition to providing a useful summary of the total data base, principal component plotting provided a valuable aid to identifying unusual sample behavior. By identifying unusual behavior investigators could focus on possible causes. (The causes for anomalous behavior are often simple measurement errors. Thus, identifying outliers helps focus attention on the distinctions that make a difference.) The summaries included pertinent graphical representations of the measurement space defined by their experimental protocol and identify the multivariate factors that describe that space. A summary of the key variable interactions was also provided from this technique.

COLLABORATION CONSIDERATIONS

We were sensitive to the investigator's desire to examine and interpret their own data. Indeed, they are uniquely capable of interpreting the results of their experimental design. However, recognizing the complexity of data interpretation when many variables are present, we view the multivariate summaries as interpretive tools to be exploited by the individual investigators.

Past experience indicated that investigator collaboration is at a maximum in the sampling and initial data gathering phases of most interdisciplinary efforts. However, in later phases individuals tend to narrow their focus to their specific experimental objectives. Collaboration often is induced only by the external stimulus of report writing deadlines. We believe that sustained immersion in into the interdisciplinary context facilitates interpretive insights. It was our desire to induce the

investigators to examine and interpret their own results in an interdisciplinary context by applying the pattern recognition techniques to different experimental measurement domains and supplying cross-disciplinary summaries. Thus, a formal information transfer mechanism was incorporated into the data analysis plan. The formal transfer was accomplished by providing periodic written data summaries called "Investigator Updates" and more frequent communication was done through letters and telephone communications. In order to preserve investigator "rights" to the data that they shared for summary purposes, no independent publications were by this researcher.

INVESTIGATOR UPDATES

As indicated earlier, insight regarding mechanistic description of the system requires illucidation of variable interactions. The need for multivariate techniques to identify these interactions is apparent when one considers that each measured parameter contributes one dimension to the representation. Thus examining two parameter interactions requires a two-dimensional plot. Such graphical representations are effective in identifying significant relationships among the variables. A three variable system requires a three dimensional plot to simultaneously represent all potential bivariate interactions. However, as the number of variables increases the dimensionality of the required representation exceeds man's ability to perceive significant patterns in the data. Indeed, humans do not conceptualize comfortably beyond three dimensions. Without assistance one would be restricted to considering only problems that are characterized by three factors. (If one restricts the interpretive task to two variable interactions one may generate a series of two-dimensional graphs, one for each unique bivariate pair. Again, the mere task of examining all of the plots becomes formidable. A data base consisting of 30 measured variables would require examining 435 plots!) One commonly computes a correlation matrix consisting of all unique bivariate correlation coefficients to summarize the variable interactions. While this type of summary is helpful, it provides little insight regarding the natural associations among groups of variables. The more powerful factor analytic treatment extracts the significant underlying relationships that characterize the data. Factor analysis provides the tools by which data are converted to information. It is in these natural associations that one hopes to find the clues to uncover otherwise obscure mechanisms.

A second capability that one needs in examining large data bases is a convenient way to represent relationships among samples or objects upon which the measurements have been made. This procedure is analogous to the search for variables that are associated with one another. Group behavior among the objects indicates that significant distinctions are possible, and the distinctions lead to useful generalizations that simplify complex systems. Preliminary data already indicate that different organisms are present in different subsurface strata. If microbial"communities" are present it is necessary to identify chemical, nutrient, and hydrologic factors that contribute to differentiation among geologic strata. The mathematical techniques employed in pattern recognition permit rapid and efficient identification of relationships and key aspects that otherwise might remain hidden in the large mass of numbers.

In addition to the quantitative difficulties attending conventional data base examination there are several qualitative limitations imposed by the nature of the measurements themselves. Many standard statistical techniques require some knowledge or assumptions about the shape of the distribution of measured values. Many environmental variables do not have well-behaved distributions; some are highly skewed and some are multimodal. Robust analysis of these data requires techniques that do not rely upon a priori knowledge or assumptions about the underlying variable distributions. Qualitative limitations and the magnitude of the interpretive task mandates a computer assisted examination of all possible relationships. We exploit the power of computer assisted pattern recognition techniques in examining the data base. Examples of how these results were communicated to investigators are provided in Appendix 1 and 2.

One of the primary objectives was to provide graphical and other multivariate summaries for investigators' use in designing new studies. Since the initial experimental protocols were designed to characterize a very complex systems they included measurement of dozens of variables. Some of the variables included in the initial experimental design were, in retrospect, redundant and could be excluded in subsequent experimental designs. Thus information was provided to investigators for modification of the Fourth experimental hole. An example of one such formal investigator communication is attached as Appendix 3.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE STUDIES

While the informal data analysis services and formal information provided through multivariate summaries in the form of Investigator Updates were useful; investigator collaboration and communication with the data analysis aspect of the work would have been better if the multivariate statistical support personnel had been brought in at the beginning of the project. Had this worker been present during the experimental design stage, I believe that investigators would have seen the integral part that a data analysis plan plays in the analysis of results. Much effort was expended in educating the investigators about the role of this aspect of the work and what collaboration was requires. Even after numerous formal requests for investigator data and extensive personal communication it was not always possible to obtain data in a timely fashion. Indeed, this and other delays in completion of the fourth investigator's hole precluded incorporation of these data into the the study prior to expiration of the contract. Future workers who may wish to undertake large data-rich interdisciplinary studies may wish to examine the recommendations regarding the design of the data-information management aspects that have been provided in Appendix 4 of this report.

APPENDIX 1

Investigator Update March 1988, Pore Water Chemistry.

SAVANNAH RIVER PROJECT
EXPLORATORY DEEP PROBE

Exploratory Data Analysis

Investigator Update

15-MARCH-1988

Pore Water Chemistry

center for environmental sciences Laboratory for Chemometrics

University of Colorado at Denver

DEEP PROBE UPDATE

From R. Meglen

Objective:

To provide a graphical supplement to Tom Garland's summary of the major cation and anion porewater chemistry data.

TRILINEAR DIAGRAMS

Geochemists often use a graphical technique (Trilinear Diagrams) developed by Piper to examine the relationships of waters that may have different origins. Waters that differ in their major ionic components plot in different regions of these diagrams and give a pictorial representation of similarity/dissimilarity among the waters. A brief description of these plots is appended for those who might be interested in more detailed explanation. I have prepared trilinear diagrams from the pore water ionic compositions provided by Tom Garland. The trilinear anion plot (Figure 1) shows that sulfate dominates anionic composition in about two thirds of the waters. The cationic trilinear plot (Figure 1) shows that about two thirds of the waters are dominated by sodium. Some exceptions to the generalization that sodium sulfate is the dominant water composition should be noted. Sample number 3 (P24-190-1, Dry Branch) is the only pore water in the calcium bicarbonate domain. Samples 5, 6, 10, 18 (See Table for sample identifiers.) are the only calcium sulfate waters. Samples 1, 2, 16, 30 are the only brine-like sodium chloride waters. Samples 4 and 17 are sodium with approximately equal content of bicarbonate and sulfate. It is interesting to note that the drilling mud and K-25 well water composition plot in this region. The fact that these two samples are well displaced from nearly all other pore waters is indirect evidence that the drill muds did not penetrate the most of the cores. If they had, one might have expected that M and W would have plotted near more of the pore waters. I have designated the top-most waters from P-24, P-28, and P-29 with plot symbols T1, T2, and T3 respectively. They contain equal amounts of calcium and sodium with a 20% contribution from magnesium. Similarly, the anionic component consists of about 60% sulfate and equal proportions of bicarbonate and halide. It is interesting that even these near surface pore waters are quite different from the other pore waters, and the mud. I think this indicates that, at least with respect to pore water, sample integrity has been maintained.

By now the investigators have read the pore water chemistry summary which appears in the draft of Science article quoted below. The two sentences concisely characterize the details of the more extensive examination provided in the trilinear plots shown here.

> "Pore water ionic composition in the formations is dominated by $CaSO_4$ in the Upland, NaCl in the Dry Branch, Na^+:$Ca^{\frac{1}{2}}$ and Cl^-:HCO_3^- in the McBean, $NaHCO_3^-$ in the Congaree, Na^+:$Ca^{\frac{1}{2}}$ sulfate in the Ellenton, Na^+:$Ca^{\frac{1}{2}}$ sulfate in the highest segments of the Upper Middendorf, and primarily NaCl in the remaining segments of the Middendorf Formation."

TABLE 1
PORE WATER CHARACTERIZATION

ID#	Sample	Formation	Cations	Anions
T1	P24-0.2	Upland	Ca/Na	SO_4:Cl
1	P24-113	Tobacco Road	Na	Cl
2	P24-147-1	Dry Branch	Na	Cl
3	P24-190-1	Dry Branch	Ca	HCO_3
4	P24-299	Congaree	Na	HCO_3:SO_4
5	P24-387	Ellenton	Ca	SO_4
6	P24-457	Ellenton	Ca	SO_4
7	P24-477	Pee Dee	Ca/Na	SO_4:Cl/HCO_3
8	P24-592	Pee Dee	Na	SO_4:HCO_3:Cl
9	P24-657	Black Creek	Na	SO_4:HCO_3
10	P24-668	Black Creek	Ca	SO_4
11	P24-777	Middendorf	Na/Ca	SO_4
12	P24-803	Middendorf	Na/Ca	SO_4
13	P24-835	Middendorf	Na/Ca	SO_4
14	P24-851	Middendorf	Na	SO_4:HCO_3
T2	P28-0.2	Tobacco Road	Ca/Na	SO_4:Cl/HCO_3
15	P28-47	Tobacco Road	Na	SO_4
16	P28-103	Dry Branch	Na	Cl
17	P28-193	Congaree	Na	Cl
18	P28-235-1	Williamsburgh	Ca	SO_4
19	P28-367	Pee Dee	Na	SO_4:HCO_3:Cl
20	P28-376	Pee Dee	Ca/Na	SO_4:Cl
21	P28-440-1	Pee Dee	Na	SO_4
22	P28-589	Middendorf	Na	SO_4
23	P28-599	Middendorf	Na	SO_4:Cl
24	P28-628	Middendorf	Na	SO_4:HCO_3
25	P28-667-1	Middendorf	Na	SO_4
26	P28-667-2	Middendorf	Na	SO_4
27	P28-670	Middendorf	Na	SO_4
28	P28-705	Middendorf	Na	SO_4:HCO_3
29	P28-709	Middendorf	Na	SO_4
T3	P29-0.2	Upland	Ca/Na	SO_4:Cl/HCO_3
30	P29-25	Tobacco Road	Na	Cl
31	P29-94	Dry Branch	Na	SO_4:Cl
32	P29-128	Congaree	Na	SO_4:HCO_3
33	P29-225	Pee Dee	Na/Ca	SO_4
34	P29-309	Pee Dee	Na	SO_4
35	P29-365	Black Creek	Na	SO_4
36	P29-463	Black Creek	Na	SO_4
37	P29-496	Middendorf	Na	SO_4
38	P29-576	Middendorf	Na	SO_4
39	P29-594	Middendorf	Na/Ca	SO_4:Cl
40	P29-612	Middendorf	Na	SO_4:Cl
41	P29-634	Middendorf	Na	SO_4
42	P29-655	Middendorf	Na	SO_4:HCO_3
43	P29-700	Middendorf	Na	SO_4:Cl
M		Drilling Mud 6/13/86	Na	HCO_3:SO_4
W		K-25 Well Water	Na/Ca	HCO_3:SO_4

Trilinear Diagram

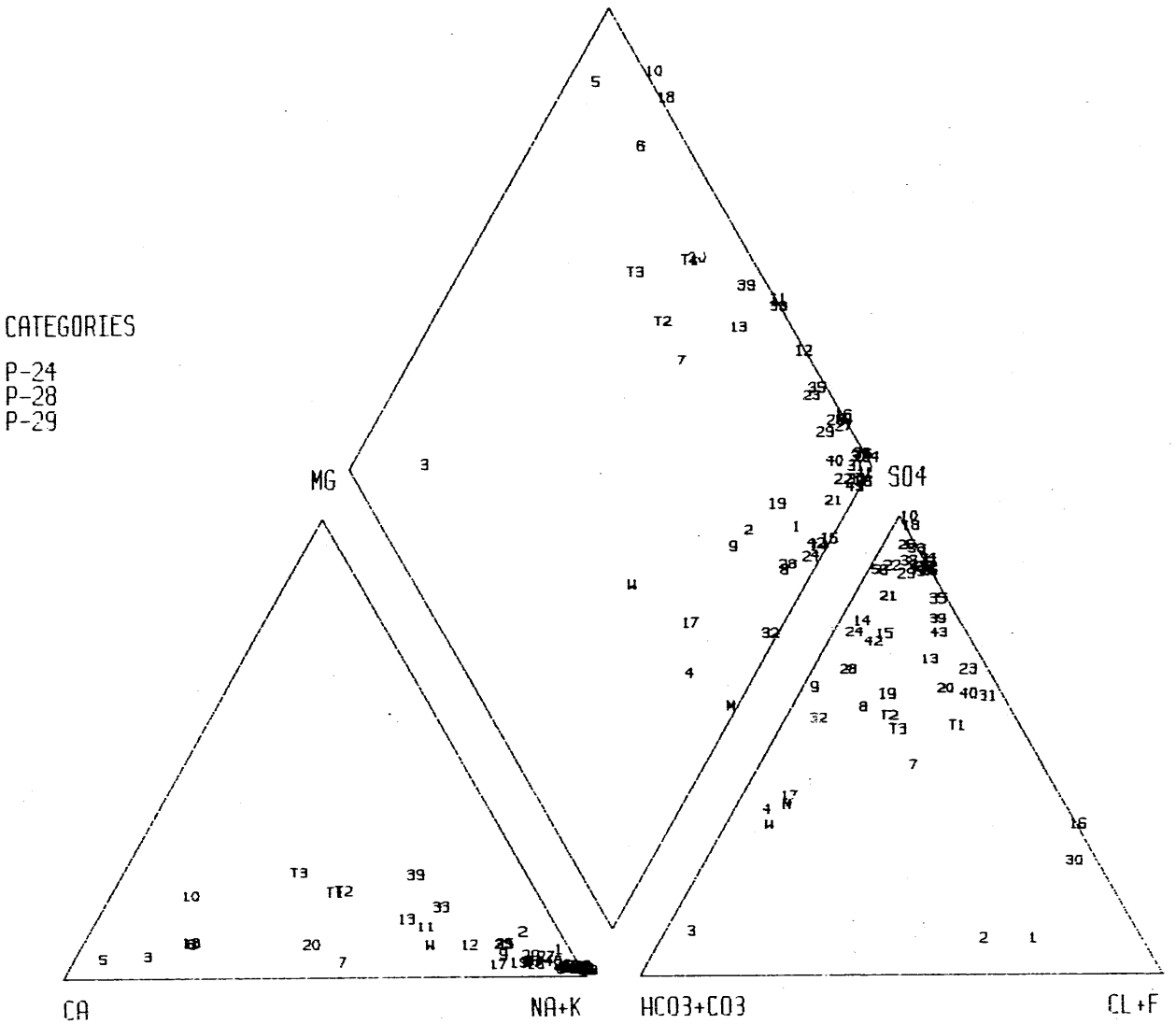


Figure 1

PRINCIPAL COMPONENT ANALYSIS

The verbal summary provided by Tom Garland is quite useful, and the graphical presentation through trilinear diagrams clearly indicates that there are major ion chemical fingerprints that characterize "groups" of pore waters. However, neither of these formats adequately characterizes the vertical variability. It is desireable to produce graphical summary that can permits vertical and spatial comparisons. Therefore, I have applied principal component analysis to the major ion pore water chemistry data in order to objectively determine the minimum number of chemical components that describe these data and vertical profiles that summarize them. The starting data base consisted of the same data that were used to prepare the trilinear plots; namely, HCO_3, Cl, F (elec), NO_3, SO_4, Ca, Mg, Na, K, NH_4. However, instead of grouping the anions and cations into chemically similar groups as in the Piper method, the principal component analysis finds the "natural" associations among the variables. The principal component analysis shows that about 75% of the variability among all of the pore waters can be accounted for by three underlying variables. The first so-called factor consists of Mg, Ca, K, and SO_4. It accounts for 47% of all the variablility among samples. That means that the chemistry factor that best distinguishes among the waters is due to these chemical constituents. This is confirmed by the spread of about one third of the samples toward the $CaSO_4$ corner of the trilinear diamond plot. We shall call this the calcium sulfate factor.

The second principal component accounting for about 18% of the variability among all pore waters consists of Na, Cl, HCO_3, and SO_4. It distinguishes sodium chloride, sodium bicarbonate and sodium sulfate pore waters from the calcium sulfate waters which lie along the Factor 1 axis. We shall call this the sodium axis. Both of these factors are consistent with our "understanding" of geochemical characteristics. That is, the observed associations among the variables which group together on these two factors are interpretable in the convensional context of waters in contact or at equilibrium with known mineral types.

The final principal component, accounting for about 11% of the total variance, consists mainly of nitrate, with moderate contributions from fluoride and ammonium ion. This is an interesting factor because it indicates that these components are acting independently of the chemical constituents in principal components one and two. It means that nitrate is uncorrelated to the other chemical constituents. This might be expected since the minerals that lead to nitrate content of the pore waters are in general quite different from the minerals that lead to the chemistries suggested by Factors 1 and 2. However, the apparent independence may also be due to an additional influence that increases the nitrate content of some samples. (Different redox characteristics or microbial solubilization, etc. ???????) Furthermore, it turns out that nitrate is inversely correlated to ammonium and fluoride. This indicates that as ammonium and fluoride content increase nitrate decreases. It may be a little dangerous to carry this interpretation too far since the inverse correlations are weak.

The purpose for performing the principal component analysis is to discover natural associations from which one may infer geochemical characteristics. But more importantly, the product of the analysis yields objective quantitative numbers (called factor or principal component scores) that permit

comparisons among the waters. And this capacity allows us to graphically examine the spatial (vertical profile) variability among all of the samples in a way that was not possible with the trilinear plots or verbal descriptions. The principal component scores indicate how each sample compares to all others. Thus, a Thus, if a sample has a score of -2.5 on principal component 1 we know that it is 2.5 standard deviations below the mean in the combined content of Mg, Ca, K, and SO_4. If the same sample has a score of +1.0 on principal component 2 we know that it is one standard deviation above the mean of all other samples. Thus, I have prepared plots of the three principal components (Figures 2-4) described above as a function of depth for the core holes.

Figure 2 depicts the scores for Factor 1 versus depth. Scores for each core are depicted on separate axes; with negative scores plotted to the left and positive scores to the right. The points are connected with line segments between successive points (sample locations) and shaded in order to enhance recognition of vertical patterns. For example, the first two samples from P-24 (depth 113 & 147) are -1.5 and -1.9 standard devoations below the mean of all samples in Factor 1 ($CaSO_4$). Sample number 10 at 668 feet is 2.9 standard deviations above all others in Factor 1. (Examination of the original data shows that this pore water sample has the highest Ca (473ppm) and SO_4 (124ppm) of all samples.)

> One note of caution when examining these plots is that when the scores traverse between positive and negative scores they produce a node which makes it look like there is a sharp demarcation between zones. This is an artifact of the density of sampling points. Thus, the exact vertical location of the transition points may not be adequately determined from the plots. However, the general character above and below a node indicates that somewhere between successive high positive and negative scores there is a major change in chemical character of the pore waters.

Examination of the three core profiles indicates that there are four major regions. In all three holes $CaSO_4$ is 1 to 2 standard deviations below the average in the Dry Branch and Tobacco Road. A transition occurs and in the Ellenton region the $CaSO_4$ content becomes 1 to 2 standard devaitions higher than the average. In between the Ellenton and Pee Dee another transition occurs and the Pee Dee pore waters are again below average in $CaSO_4$. Only P-24 shows a clear $CaSO_4$ rich region in the Black Creek. Cores P-28 and P-29 in this region are quite close to average.

Examination of the Factor 2 depth profile (Figure 3) for the sodium factor is more complex. In overview P-24, P-28, and P-29 are not as simliar as they are in $CaSO_4$. However, all three show the Congaree is above average in NaCl and they undergo a transition to below average in the Ellenton. Below the Pee Dee only a couple pore waters (in P-29) are more than 1 standard deviation above the average NaCl content. It is interesting to superimpose the Factor 1 and Factor 2 profiles. What you will observe is that the two factors are orthogonal; i.e. they are independent chemical behaviors. When $CaSO_4$ is very high or low NaCl tends to be the opposite.

Examination of the third factor profile (Figure 4) indicates that in contrast to pore water similarities among the holes seen for Factors 1 and 2, pore water nitrate character of P-29 is very different from P-24 and P-28. All of the pore water samples in P-24 and P-28 are below average in

nitrate, and all of the P-29 nitrates are above average. This indicates a bimodal distribution, i.e. nitrate concentration is either high or low. And only P-29 seems to have any nitrate. (Note that the sign of the nitrate contribution to the factor is negative. Thus, the highest nitrate, lowest fluoride and ammonium are to the left.) It will be interesting to examine the nitrate profile of the fourth hole. Is the difference in nitrate profiles due to some spatial geochemical anomally, a possible nitrate contamination, a spatial microbially related inhomogeneity, etc?

I realize that many of you may be unfamiliar with viewing these multivariate plots. It is a bit confusing at times. But when examining such a large number of simultaneously varying parameters we must rely on the multivariate summary techniques. In this task, the gouping of variables by similarity for graphical display is the only manageable approach. I have extensively examined the pore water chemistry in much greater detail than given here, but for now I hope that this update will help you to understand the nature of the information present in the major ion pore water data. In any case, as I proceed with the exploratory analysis of the whole data base, I will be relying on similar plots. The plots provided here will help to familiarize you with their interpretation. The next update will examine 117 chemical, well log, and microbiological variables.

SAVANNAH RIVER EXPLORATORY DEEP PROBE

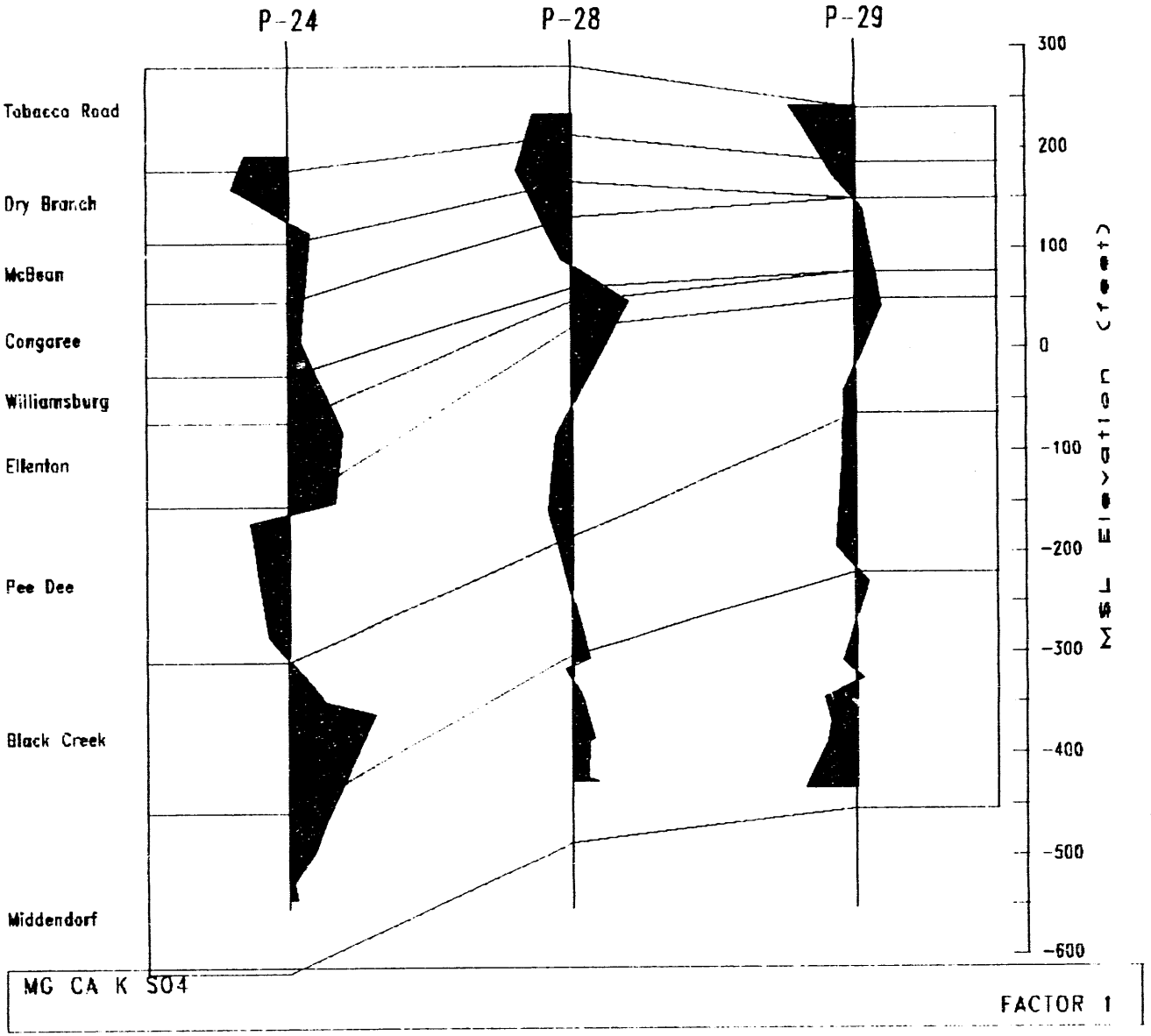


Figure 2

SAVANNAH RIVER EXPLORATORY DEEP PROBE

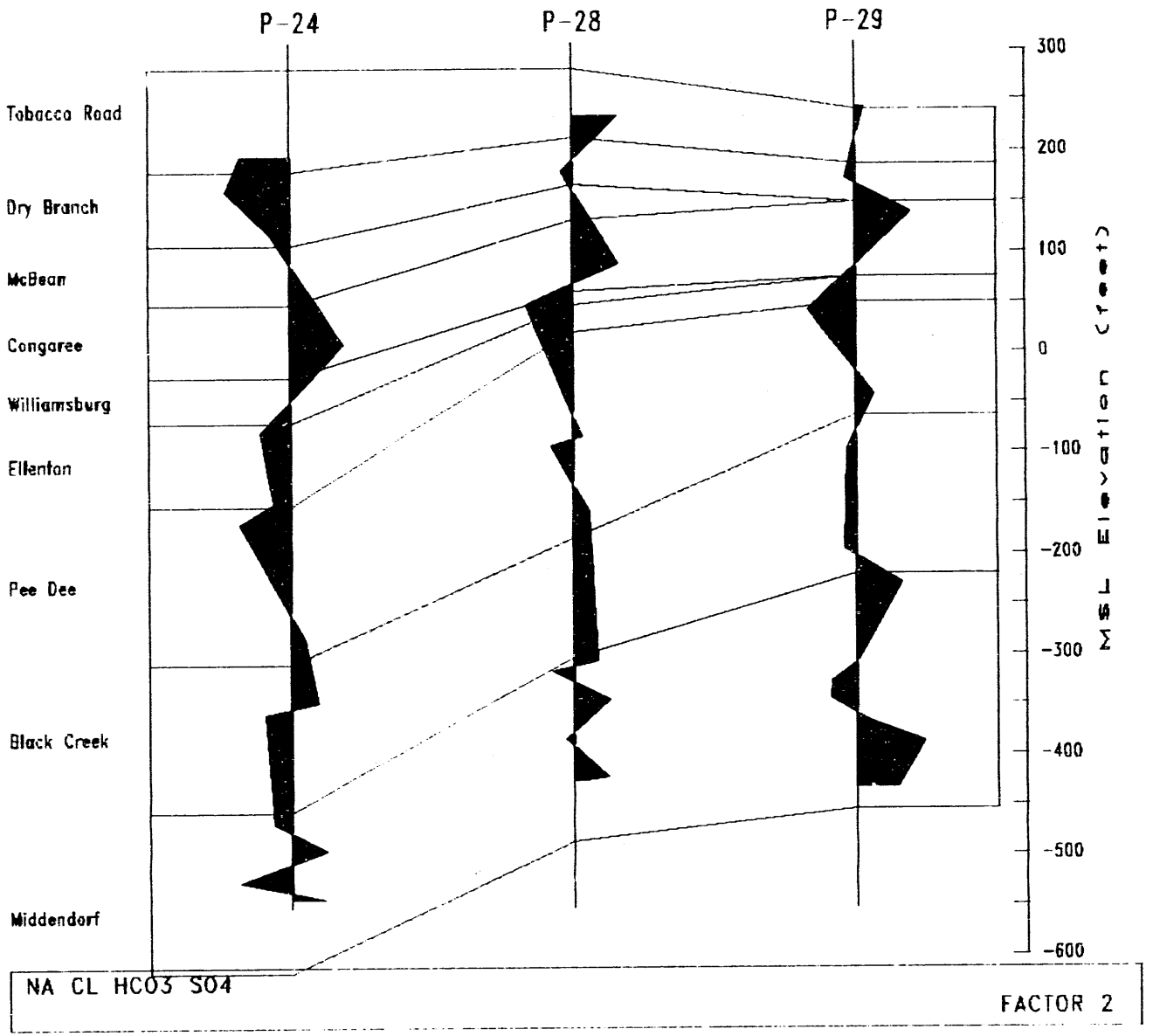


Figure 3

SAVANNAH RIVER EXPLORATORY DEEP PROBE

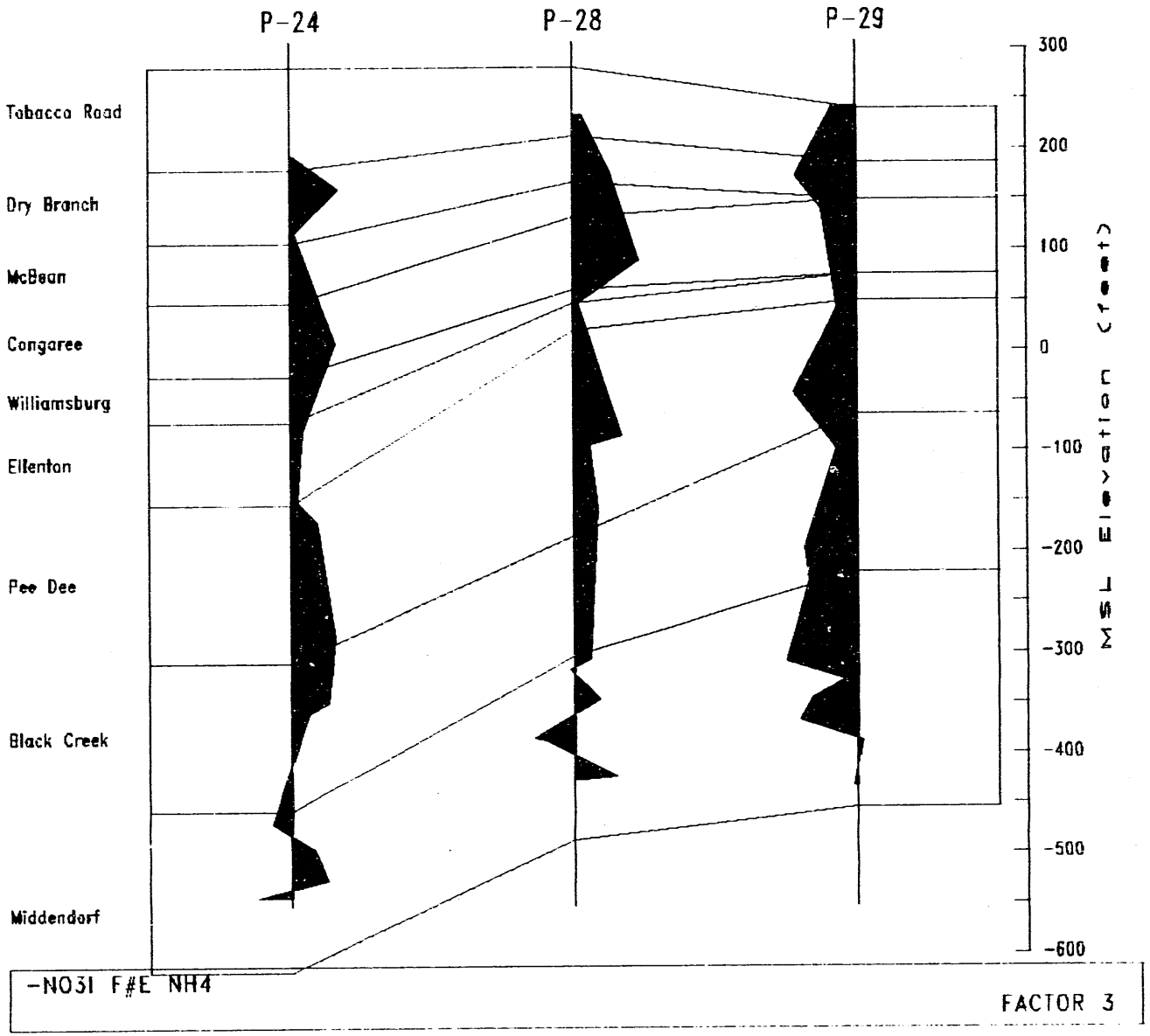


Figure 4

APPENDIX

Piper Trilinear Diagrams

Geochemists have long recognized the difficulty in simultaneously depicting variations among several water quality parameters. An intersesting and useful graphical display technique developed by Piper (See reference) has received wide use among geochemists. These plots permit the simultaneous display of six chemical variables. The major aqueous cations are divided into three contributions (Ca, Mg, & $Na+K+NH_4$) computed in chemical equivalents. The major anions are also divided into three contributions (SO_4, $Cl+F+NO_3$,CO_3+HCO_3) computed in chemical equivalent units. The data are plotted on two triangular fields. The percent contribution from each of the three cation groups is plotted in the left triangular field according to conventional trilinear coordinates. Thus a single point is described by three coordinates (Ca, Mg, $Na+K+NH_4$). A corresponding point for the anionic content of that water sample is plotted on the right triangular field. A point near a vertex indicates the dominance of that component. A point near the center of the triangular field indicates that all components contribute equally to the chemical character of the water. Thus, in Figure 1, point number 1 indicates a water whose cationic character is 1/3 Ca, 1/3 Mg, and 1/3 in alkali metals. Similarly, the anionic composition of this water is 1/3 carbonate, 1/3 sulfate, and 1/3 halide. (For geochemical reasons nitrate is included with the halides.) The cation plot shows that water number 2 is mainly Ca and the anion composition is dominated by sulfate. Similarly, water number 3 is a magnesium halide water. A point half way between two vertices and near the margin indicates a water which has no contribution from the opposing vertex's component. Thus, point 5, on the cation plot indicates that the water contains a 50%/50% mixture of Ca and the alkali metals. This position of point 5 on the anion plot indicates that the anionic composition of the water is mainly halides.

These trilinear diagrams are useful in comparing waters that have different major ion characteristics. The comparison among waters may be simplified further by creating an additional diamond shaped field that graphically combines the data points from the two triangular fields. This is shown in Figure 2. The location of a point in the diamond field is obtained by projecting the cation and corresponding anion points along the dimension parallel to the diamond's sides. Thus, each water is depicted as a single point in the diamond field. This plot provides a six dimensional representation of major chemical components of the waters.
Figure 3 summarizes the major regions that characterize waters of various geochemical types. Geochemical generalizations regarding various water types are given.

Reference: Piper, A. M., "A Graphic Procedure in the Geochemical Interpretation of Water Analyses", *Groundwater Notes in Geochemistry*, #12, pp1-14(**1952**).

Piper Trilinear Diagrams

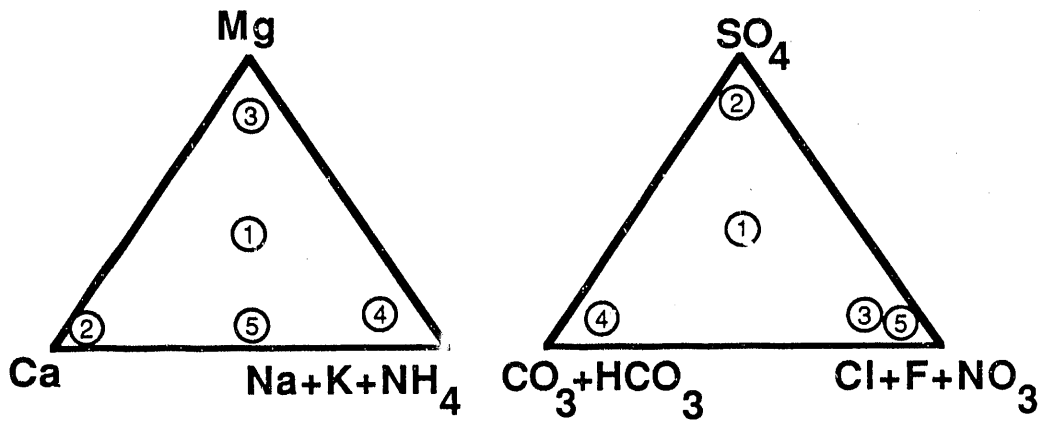


Figure 1

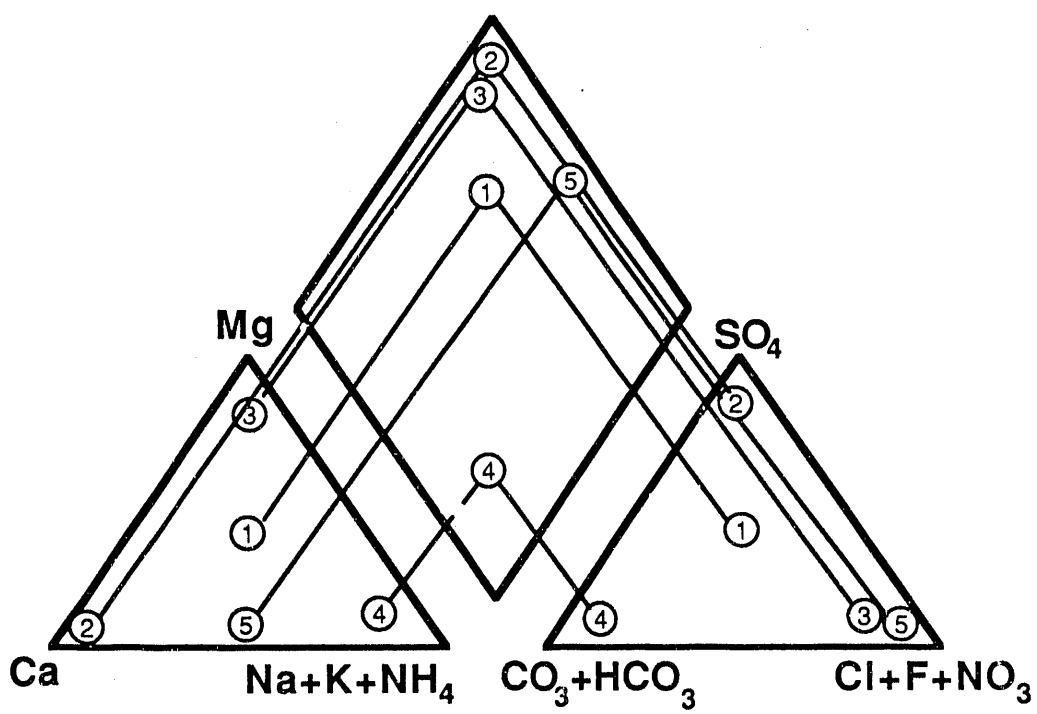


Figure 2

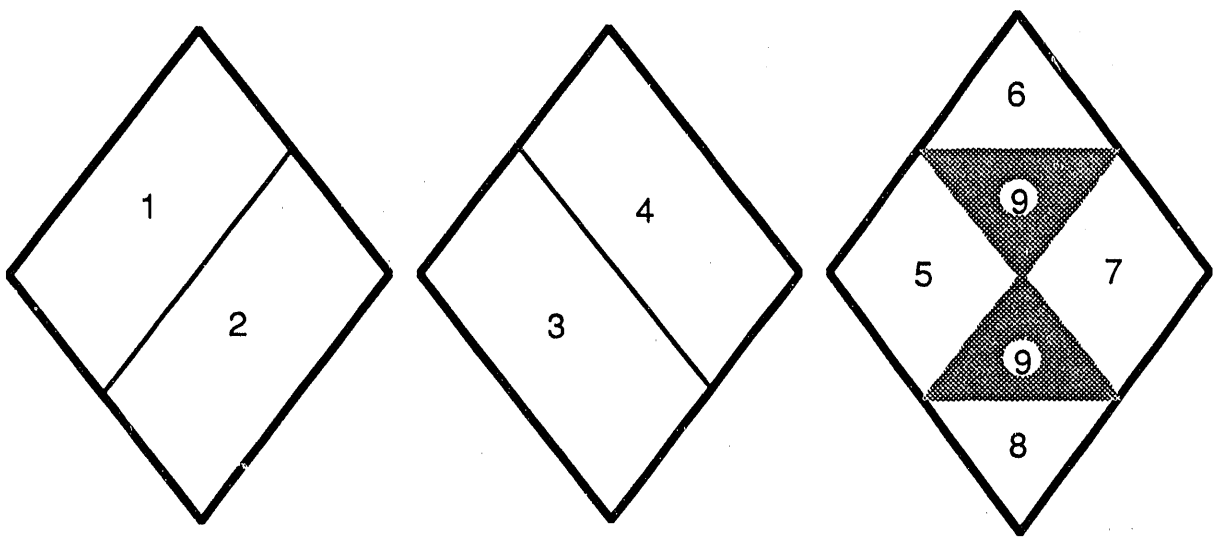


Figure 3

Subdivisions of the diamond-shaped field

1. Alkaline earths exceed alkalis
2. Alkalis exceed alkaline earths
3. Weak acids exceed strong acids
4. Strong acids exceed weak acids
5. Carbonate hardness ("Secondary alkalinity") exceeds 50%, i.e., chemical properties of the water are dominated by alkaline earths and weak acids.
6. Noncarbonate hardness ("secondary salinity") exceeds 50%
7. Noncarbonate alkai ("primary salinity") exceeds 50%, i.e., chemical properties are dominated by alkalies and strong acids. Ocean water and many brines plot in this area, near its right-hand vertex.
8. Carbonate alkali ("primary alkalinity") exceeds 50%. Waters which are inordinately soft in proportion to their content of dissolved salts plot in this region.
9. No one cation-anion pair exceeds 50%.

APPENDIX 2

Investigator Update April 1988, Correlations Among Variables and Principal Component Analysis.

SAVANNAH RIVER PROJECT
EXPLORATORY DEEP PROBE

Exploratory Data Analysis

Investigator Update

15-APRIL-1988

Correlations Among Variables
&
Principal Component Analysis

center for environmental sciences Laboratory for Chemometrics

University of Colorado at Denver

DEEP PROBE UPDATE

From R. Meglen

Objective:

To provide a summary of statistical correlations found among the microbiology, pore water chemistry, and geological variables from three cores P-24, P-28, P-29.

One of the tasks that I have undertaken is to provide investigators with assistance in examining their data in the larger context of data obtained by all investigators. The objective is to find significant relationships among the chemical, geological and microbiological variables. Since the number of variables is large it is necessary to use multivariate statistical techniques to summarize these relationships. Therefore, this Update will describe my initial examination of the data that I have received so far.

SUMMARY

The matrix of correlations among 110 variables has more than 400 very highly (99%) significant correlations with r greater than +/- 0.6. (Figure 1.)

Scattergrams showing 134 bivariate pairs correlated at greater than r = +/- 0.7 are provided (Figures 2-9).

Principal component analysis of the correlation matrix reveals that nine factors account for 72% of the total variance in the data base.

Five factors suggest five independent microbiological sources of variablility among the 43 samples.

Two geochemical factors are revealed. One related to variability in the major ion chemical charactersitics in pore waters; the other identifies unique nitrate behavior in core P-29.

Profile plots showing the vertical variability of the factors in each of the cores are provided (Figures 12-20). (An alternate method of display is provided in Figures 21-29.)

Suggestions about how investigators may wish to use the information provided in this summary are also given.

Investigators are requested to examine the list of variables and to provide additional data that may have been acquired since the data used in this summary were obtained.

DATA BASE

Not all investigators have supplied me directly with their data. Much of what I have was forwarded to me by Carl Fliermans during his preparation of the publication and meeting presentation. Therefore, I am not sure of who should be credited for the measurements in the data that I used in this first pass at a comprehensive summary. Some of the data sent to me is not complete for all three holes. Therefore, the only variables that I included in this data summary are ones for which I have data for all three holes. While data are available for the top-most sample (Depth 0.2 ft.) for each core, I have deleted them from consideration because they represent influential points on almost all variables. (Statistically, "influential" points are points that lie at measurement extrema and are significantly distant from the more general distribution of measurements.) Strictly speaking they are NOT outliers in the traditional sense, but since they tend to dominate behavior and obscure the representation of the the subsurface I have held them in abeyance. Table 1 shows the samples used in the analysis.

FORTY-THREE (43) samples at depths greater than 0.2 ft. were used for all studies that are described here.

Measurements of 120 variables on these 43 samples were examined for their completeness and suitability to statistical treatment. Several porewater chemistry variables and a few others showed that measureable quantities above detection limit were present in only a few samples. Thus, they were deleted from further consideration. Table 2 shows the list of variables used in the data analysis. Descriptions of the variables and the variable abbreviations are also provided in the table.

ONE HUNDRED AND TEN (110) variables were used in the study.

TABLE 1
SAMPLES USED FOR CORRELATION AN PRINCIPAL COMPONENT ANAYSIS

ID#	Sample	Formation	
T1	P24-0.2	Upland	**********NOT USED********
1	P24-113	Tobacco Road	
2	P24-147-1	Dry Branch	
3	P24-190-1	Dry Branch	
4	P24-299	Congaree	
5	P24-387	Ellenton	
6	P24-457	Ellenton	
7	P24-477	Pee Dee	
8	P24-592	Pee Dee	
9	P24-657	Black Creek	
10	P24-668	Black Creek	
11	P24-777	Middendorf	
12	P24-803	Middendorf	
13	P24-835	Middendorf	
14	P24-851	Middendorf	
T2	P28-0.2	Tobacco Road	**********NOT USED********
15	P28-47	Tobacco Road	
16	P28-103	Dry Branch	
17	P28-193	Congaree	
18	P28-235-1	Williamsburgh	
19	P28-367	Pee Dee	
20	P28-376	Pee Dee	
21	P28-440-1	Pee Dee	
22	P28-589	Middendorf	
23	P28-599	Middendorf	
24	P28-628	Middendorf	
25	P28-667-1	Middendorf	
26	P28-667-2	Middendorf	
27	P28-670	Middendorf	
28	P28-705	Middendorf	
29	P28-709	Middendorf	
T3	P29-0.2	Upland	**********NOT USED********
30	P29-25	Tobacco Road	
31	P29-94	Dry Branch	
32	P29-128	Congaree	
33	P29-225	Pee Dee	
34	P29-309	Pee Dee	
35	P29-365	Black Creek	
36	P29-463	Black Creek	
37	P29-496	Middendorf	
38	P29-576	Middendorf	
39	P29-594	Middendorf	
40	P29-612	Middendorf	
41	P29-634	Middendorf	
42	P29-655	Middendorf	
43	P29-700	Middendorf	

TABLE 2

VARIABLE	UNITS	TRANSFORM	VARIABLE ABBREV.	
Vol in Frac.	mL		VOL	
Initial Moist	%		INIM	
Incub. Moist	%		INCM	
Conductivity	uS/cm	Log	COND	
pH			PH	
Eh			EH	
Diss. Org. Carbon	ug/mL	Log	DOC	
Diss. Inorg. Carbon	ug/mL	Log	DIC	
NH4	ug/mL	Log	NH4	
NO3 (color)	ug/mL	Log	NO3C	
NO3 (Ion Chrom.)	ug/mL	Log	NO3I	
NO2 (color)	ug/mL	Log	NO2C	
NO2 (Ion Chrom.)	ug/mL	Log	NO2I	
F (Ion Sel. Elec.)	ug/mL	Log	F#E	
F (Ion Chrom.)	ug/mL	Log	F#I	
Cl (Ion Chrom.)	ug/mL	Log	CL	
Br (Ion Chrom.)	ug/mL	Log	BR	
PO4	ug/mL	Log	PO4	
SO4	ug/mL	Log	SO4	
S2O3	ug/mL		S2O3	*******NOT USED*******
Fe+3	ug/mL	Log	FE3	
Fe+2	ug/mL	Log	FE2	
Al	ug/mL	Log	AL	
B	ug/mL	Log	B	
Ba	ug/mL	Log	BA	
Ca	ug/mL	Log	CA	
Cd	ug/mL	Log	CD	
Co	ug/mL	Log	CO	
Cr	ug/mL	Log	CR	
Cu	ug/mL	Log	CU	
Fe (Tot.)	ug/mL	Log	FE	
K	ug/mL	Log	K	
Li	ug/mL	Log	LI	
Mg	ug/mL	Log	MG	
Mn	ug/mL	Log	MN	
Mo	ug/mL	Log	MO	
Na	ug/mL	Log	NA	
Ni	ug/mL	Log	NI	
P (Tot.)	ug/mL		P	*******NOT USED*******
Pb	ug/mL		PB	*******NOT USED*******
Sb	ug/mL		SB	*******NOT USED*******
Si	ug/mL	Log	SI	
Sr	ug/mL	Log	SR	
Te	ug/mL		TE	*******NOT USED*******
Ti	ug/mL		TI	*******NOT USED*******
Zn	ug/mL	Log	ZN	
Anion-cation balance	uequiv		ACB	*******NOT USED*******

TABLE 2 (continued)

VARIABLE		VARIABLE ABBREV.
Viable plt. cnts. Brain heart infusion agar @ 23 C		BHI23
Viable plt. cnts. Brain heart infusion agar @ 37 C		BHI37
Viable plt. cnts. Brain heart infusion agar @ 37 C (heat shock)		BHI37H
Viable plt. cnts. Peptone-tryptone-yeast extr.-glucose agar @ 4 C		PYG4
Viable plt. cnts. Peptone-tryptone-yeast extr.-glucose agar @ 23 C		PYG23
Viable plt. cnts. Peptone-tryptone-yeast extr.-glucose agar @ 23 C (heat shock)		PYG23H
Viable plt. cnts. Peptone-tryptone-yeast extr.-glucose agar @ 55 C		PYG55
Viable plt. cnts. Peptone-tryptone-yeast extr.-glucose agar (1:100 diln.) @ 4 C		P14
Viable plt. cnts. Peptone-tryptone-yeast extr.-glucose agar (1:100 diln.) @ 23 C		P123
Viable plt. cnts. Peptone-tryptone-yeast extr.-glucose agar (1:100 diln.) @ 55 C		P155
Viable plt. cnts. Bacto-agar and distilled water @ 23 C		AW23
MPN Enumer. Brain heart infusion broth @ 23 C		MBHI23
MPN Enumer. Brain heart infusion broth @ 37 C		MBHI37
MPN Enumer. Brain heart infusion broth @ 55 C		MBHI55
MPN Enumer. Peptone-tryptone-yeast extr.-glucose broth @ 23 C		MPYG23
MPN Enumer. Peptone-tryptone-yeast extr.-glucose broth @ 55 C		MPYG55
MPN Enumer. Peptone-tryptone-yeast extr.-glucose broth (1:100 diln.) @ 23 C		MP123
MPN Enumer. Peptone-tryptone-yeast extr.-glucose broth (1:100 diln.) @ 55 C		MP155
MPN Enumer. Mineral-salts with citrate @ 23 C		XMSC
MPN Enumer. Mineral-salts with acetate @ 23 C		XMSA
MPN Enumer. Mineral-salts with sucrose @ 23 C		XMSS
MPN Enumer. Mineral-salts with lactose @ 23 C		XMSL
MPN Enumer. Mineral-salts with glucose @ 23 C		XMSG
MPN Enumer. Mineral-salts with mannitol @ 23 C		XMSM
Colony diversity. Brain heart infusion agar @ 37 C		CB37
Colony diversity. Brain heart infusion agar @ 37 C (heat shock)		CB37H
Colony diversity. Peptone-tryptone-yeast extr.-glucose agar @ 23 C		CP23
Colony diversity. Peptone-tryptone-yeast extr.-glucose agar @ 23 C (heat shock)		CP23H
Colony diversity. Peptone-tryptone-yeast extr.-glucose agar @ 55 C		CP55
Colony diversity. Peptone-tryptone-yeast extr.-glucose agar (1:100 diln.) @ 23 C		C123
Colony diversity. Peptone-tryptone-yeast extr.-glucose agar (1:100 diln.) @ 55 C		C155
Sand content. % international conv.		SAND
Clay content. % international conv.	CLAY	
Colony Forming Units.	(Ghiorse) per g.dw.	CFU
Acridine Orange Direct Count.	(Ghiorse) per g.dw.	AODC
MPN protozoa.	(Ghiorse) per g.dw.	PROT
Algae.	(Ghiorse) per g.dw.	ALG
Fungi.	(Ghiorse) per g.dw.	FUNGG
ATP.	(Ghiorse) per g.dw.	ATP
Total Most Probable No.	(Phelps)	TMPN
Total Plate	(Phelps) CFU	TPC
Fungal Plate	(Phelps) CFU	FUNGP
Morphologies seen	(Phelps)	MORPH
Proteolytic	(Phelps) CFU	PPC
Dilute TYEG	(Phelps)	TYEG
Saccharolytic	(Phelps)	SACC
Psychrophylic	(Phelps)	PSYC
Oligotrophic	(Phelps)	OLIG
Methanol Utilization	(Phelps)	MEOH
Methane Utilization	(Phelps)	CH4
Propane Utilization	(Phelps)	C3H8
Hydrogen Utilization	(Phelps)	HYDR

Phosphatidyl Choline	(Phelps)		PHOS
Methyl Palmitate	(Phelps)		MPAL
Catechol	(Phelps)		CAT
Halophilic	(Phelps)		HALO
Coliforms	(Suflita)		COLI
Fungi	(Suflita)		FUNGS
P-Lipids fatty acids.	(White)	pmols/g	PLIPS
C-14 Acetate Lipids	(White)	dpm/day	CALIPS
Tritiated Thymidine in DNA	(White)	dpm/day	HTHY

GEOLOGICAL CHARACTERIZATION DATA FROM WELL LOGS

VARIABLE	UNITS	VARIABLE ABBREV.
Gravel	%	GGRAV
Sand	%	GSAND
MUD	%	GMUD
Muscovite	%	GMUSC
Glauconite	%	GGLAU
Lignite	%	GLIGN
Sulfides	%	GSULF
N Gamma	cps	NGAM
S. P.	mv	SP16
Resistivity	Ohm-M	RS64
S. P. Res.	Ohms	SPRS

DATA PREPROCESSING

Statistical examination included obtaining frequency distribution plots for all variables. Cumulative frequency plots indicated that all pore water chemistry variables were nearly log normal and were therefore log transformed prior to other treatment. (Table 2 shows which variables were log transformed.) Logarithmic "units" were also used for all microbiological variables such as plate counts, MPN, etc. In cases were zero counts were reported 1.0 was added to all data before log transforming. When no measurement was available for a particular variable on a sample, the "missing" value was assigned the average value for that variable. (I have studied the "missing value" problem for several years and experience has shown that this procedure introduces the least bias to the exploratory analysis. I will be able to say more about the details in a formal way at a later date when it becomes important for final interpretation.)

CORRELATIONS

The first step in any exploratory data analysis is to search for relationships among the measured variables. While certain hypotheses might suggest that a particular pair of variables might be related, not all relationships can be anticipated. When the number of variables is large the number of bivariate correlations to compute and scatter plots to examine gets very large. For 110 variables one would have to examine 5995 (110X109/2) scatter plots and compute that number correlation coefficients. I have computed the correlation coefficients between all pairs of variables for the data base. The table of all possible correlation coefficients is 17 pages long! Instead of reproducing it here I have prepared a graphical depiction of this correlation matrix shown in Figure 1. Each correlation coefficient is depicted by a square picture element. The identity of the variable pairs may be determined by locating the variable name along the left vertical and the desired covariant along the horizontal. The magnitude of the correlation coefficient is color coded; red indicates that the correlation coefficient is greater than or equal to 0.9, magenta indicates that r is greater than or equal to 0.8 and less than 0.9, etc. (See color key on Figure 1.) Note that positive correlations are depicted as filled boxes and inverse, or negative correlations bear the same color coding but are depicted as unfilled squares. A couple of examples might help familiarize you with the use of the figure. Log of pore water sulfate (SO4) and log of pore water conductivity (COND) are correlated at greater than 0.9. Percent sand (SAND) and initial moisture (INIM) are inversely correlated with r between -0.6 and -0.7. (In case you have forgotten how to interpret these correlations I have provided a short refresher in Appendix A.)

Before dealing with specifics a few generalizations will be helpful. All of the correlations depicted are significant at greater than the 99% confidence level. (See Appendix A for explanation.) The upper left corner contains the pore water chemistry variables. As might be expected, there are several strong relationships among the pore water chemistry variables. There are no correlations greater than 0.8 between pore water chemistry variables and any other variables. A few moderate inverse correlations exist between some pore water chemistry variables and the MPN enumerations on carbon source ammended media. (More about this later.) As might be expected, there is a large number of moderate to high correlations among the enumerations and colony diversity variables with different media and temperatures.

SCATTER PLOTS

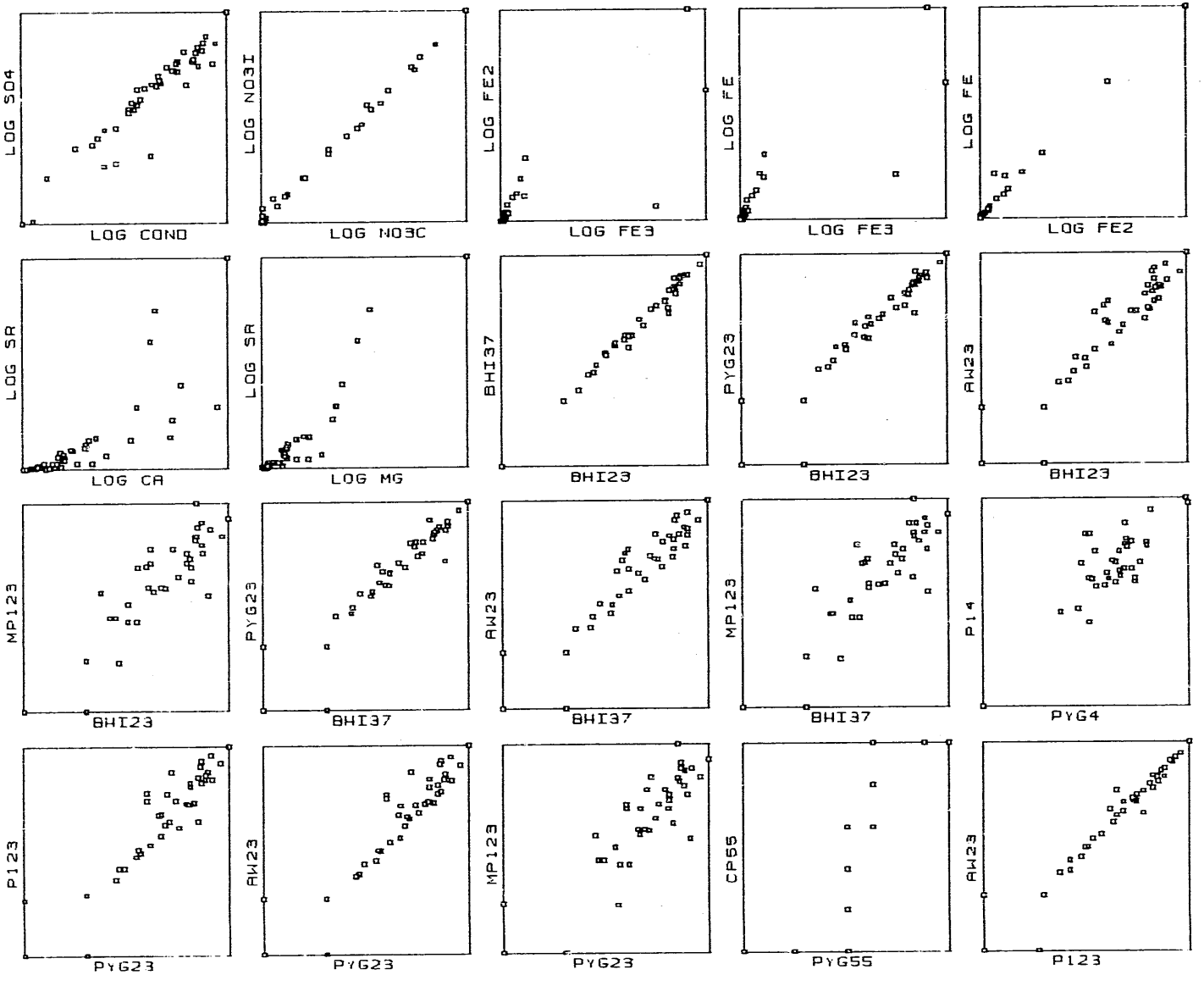
There are 134 correlations of magnitude +/-0.7 or greater. Since many individuals may find it more informative to look at scatter plots I have prepared miniature scatter plots (Figures 2-9). They are grouped by magnitude; +/-0.9, +/-0.8, +/-0.7. Relatively few of the scatter plots show that the correlation is dominated by influential points.

CORRELATION MATRIX
NEGATIVE
FILE-
COAD.MAT
VAR. LIST-
VAR.LST

SELECTED SCATTER PLOTS

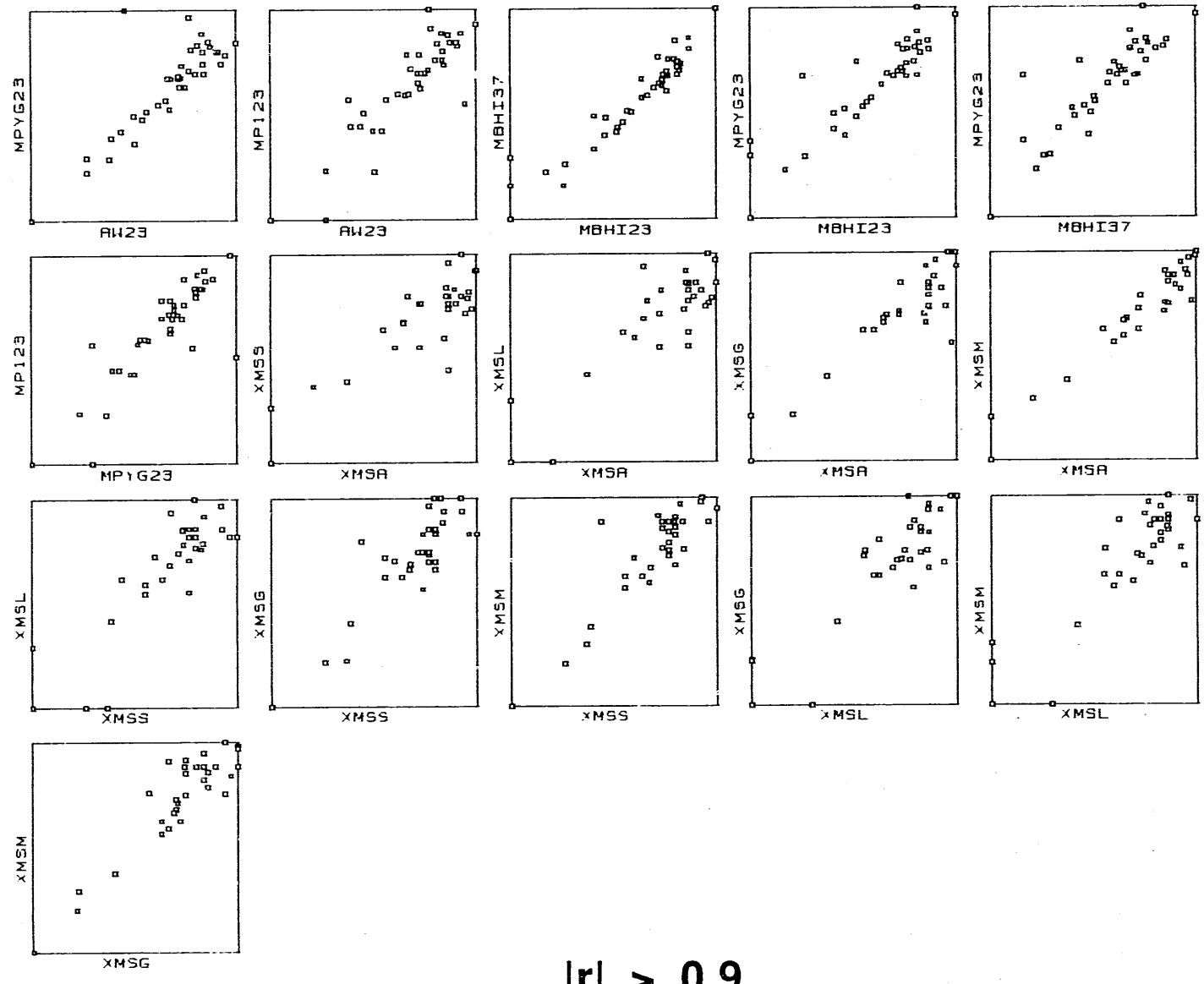
Figure 2 - 3 Correlations greater than +/- 0.9
Figure 4 - 5 Correlations between +/- 0.8 and +/- 0.9
Figure 6 - 9 Correlations between +/- 0.7 and +/- 0.8

FIGURE 2


LOG SO4
LOG COND
LOG NO3I
LOG NO3C
LOG FE2
LOG FE3
LOG FE
LOG FE3
LOG FE
LOG FE2
LOG SR
LOG CA
LOG SR
LOG MG
BHI37
BHI23
PYG23
BHI23
AW23
BHI23
MP123
BHI23
PYG23
BHI37
AW23
BHI37
MP123
BHI37
P14
PYG4
P123
PYG23
AW23
PYG23
MP123
PYG23
CP55
PYG55
AW23
P123

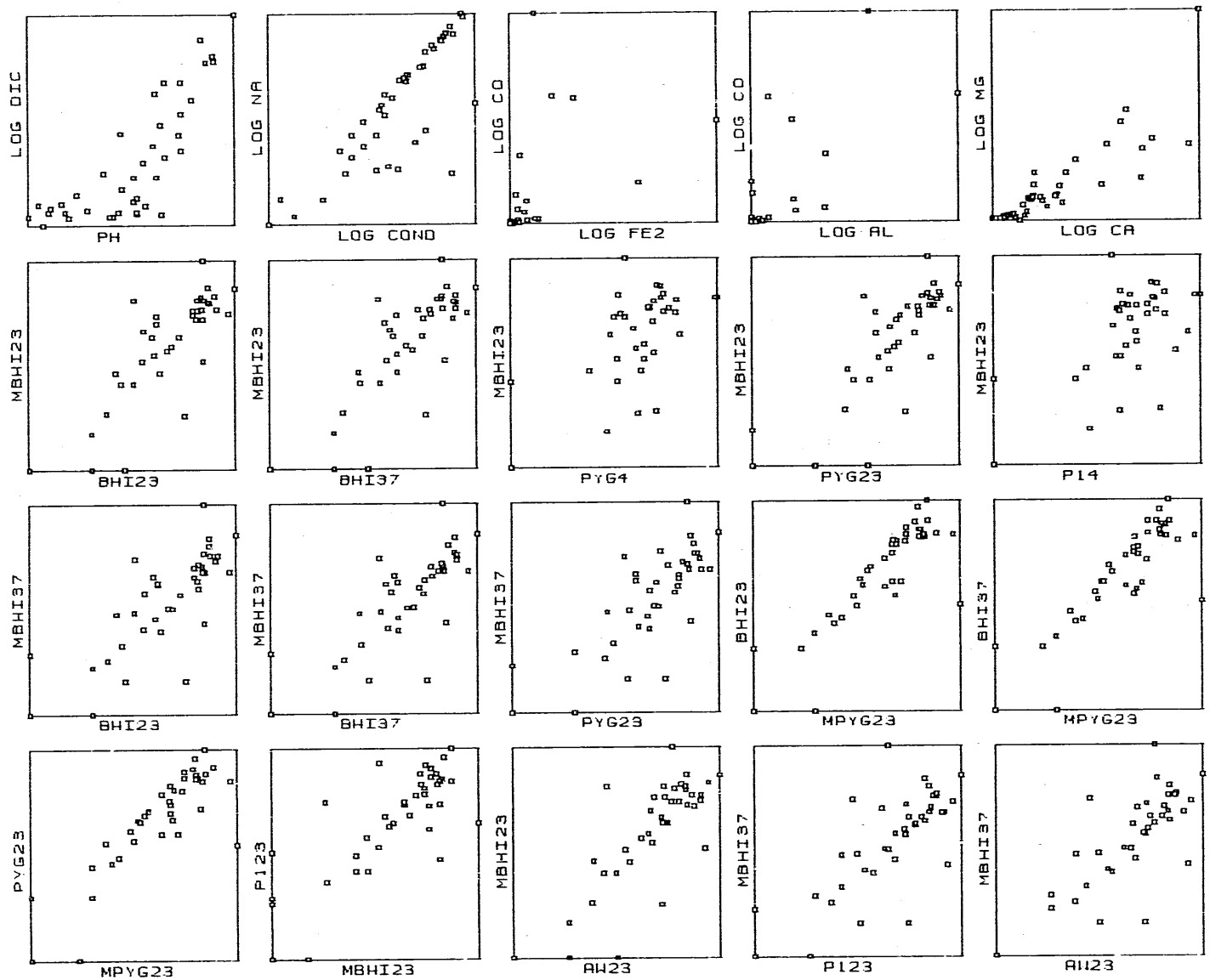
$|r| > 0.9$

FIGURE 3


MPYG23
AW23
MP123
AW23
MBHI37
MBHI23
MPYG23
MBHI23
MPYG23
MBHI37
MP123
MPYG23
XMSS
XMSA
XMSL
XMSA
XMSG
XMSA
XMSM
XMSA
XMSL
XMSS
XMSG
XMSS
XMSM
XMSS
XMSG
XMSL
XMSM
XMSL
XMSM
XMSG

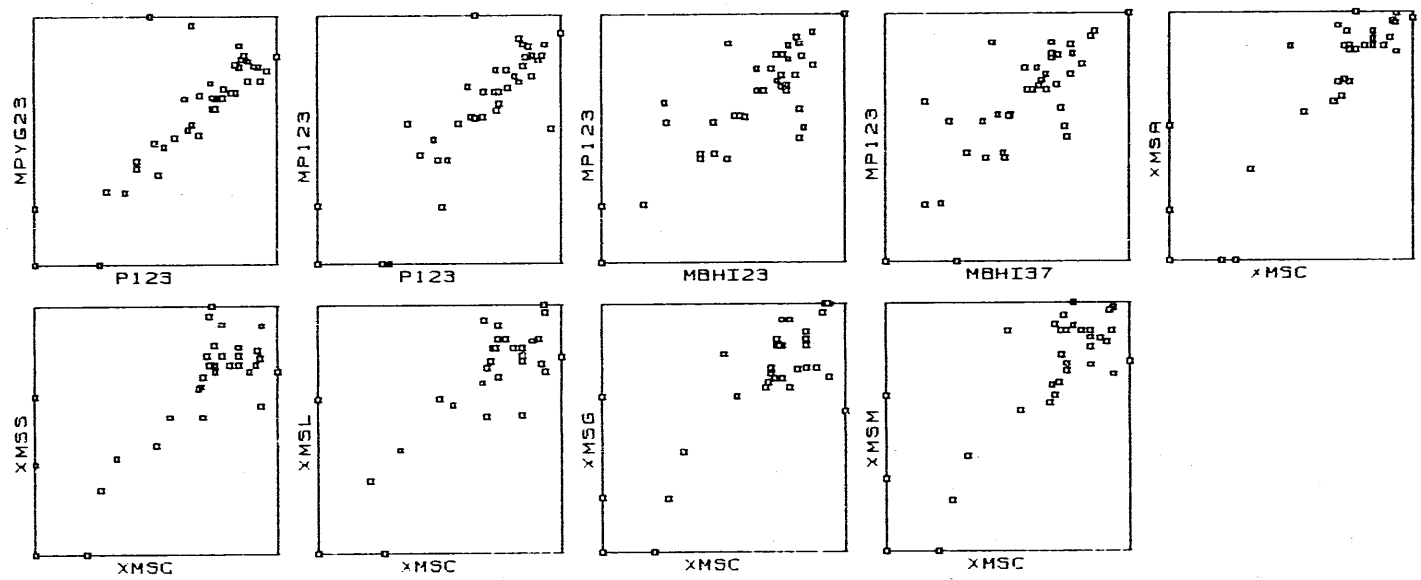
$|r| > 0.9$

FIGURE 4



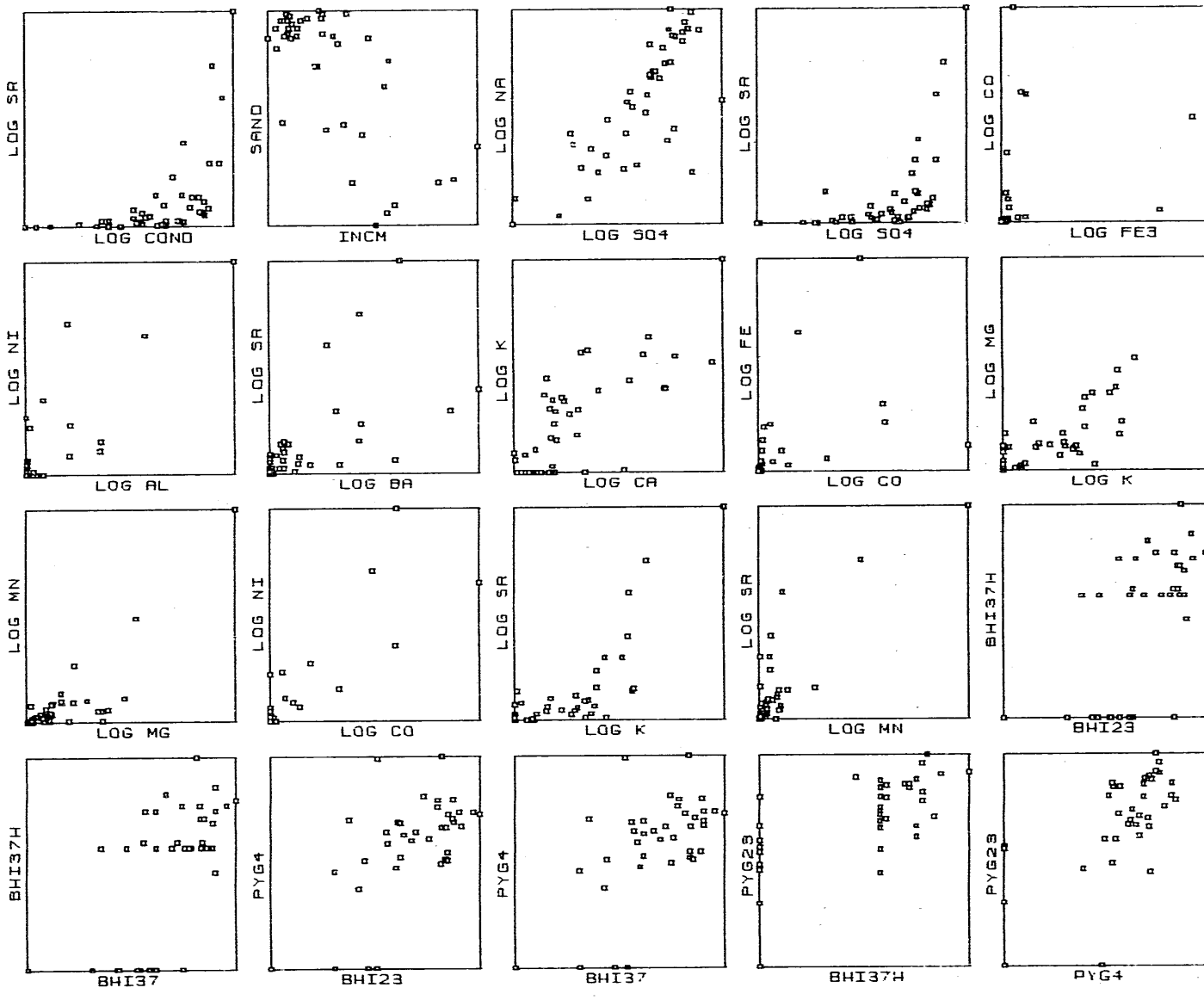
0.9 > |r| > 0.8

FIGURE 5



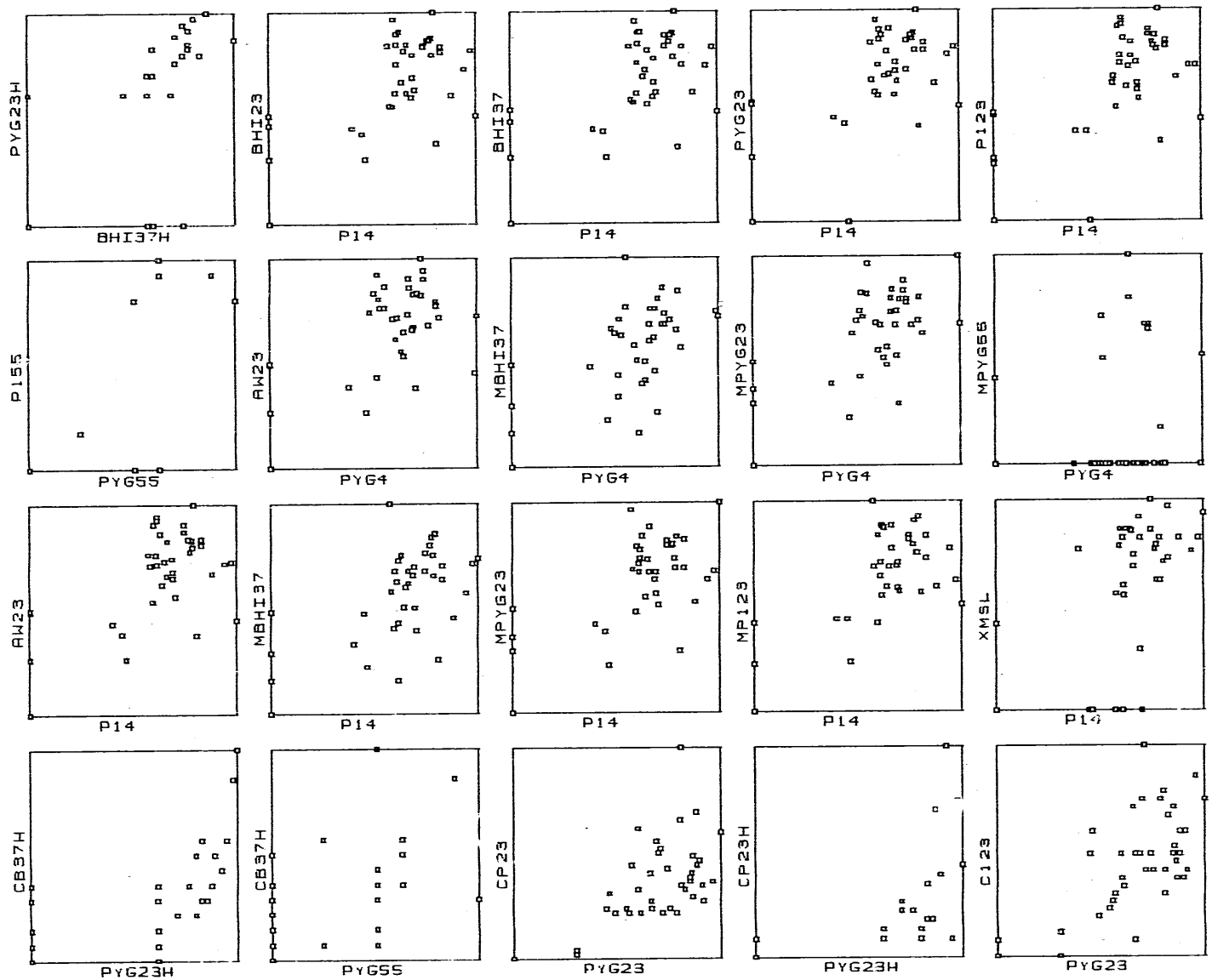
$0.9 > |r| > 0.8$

FIGURE 6



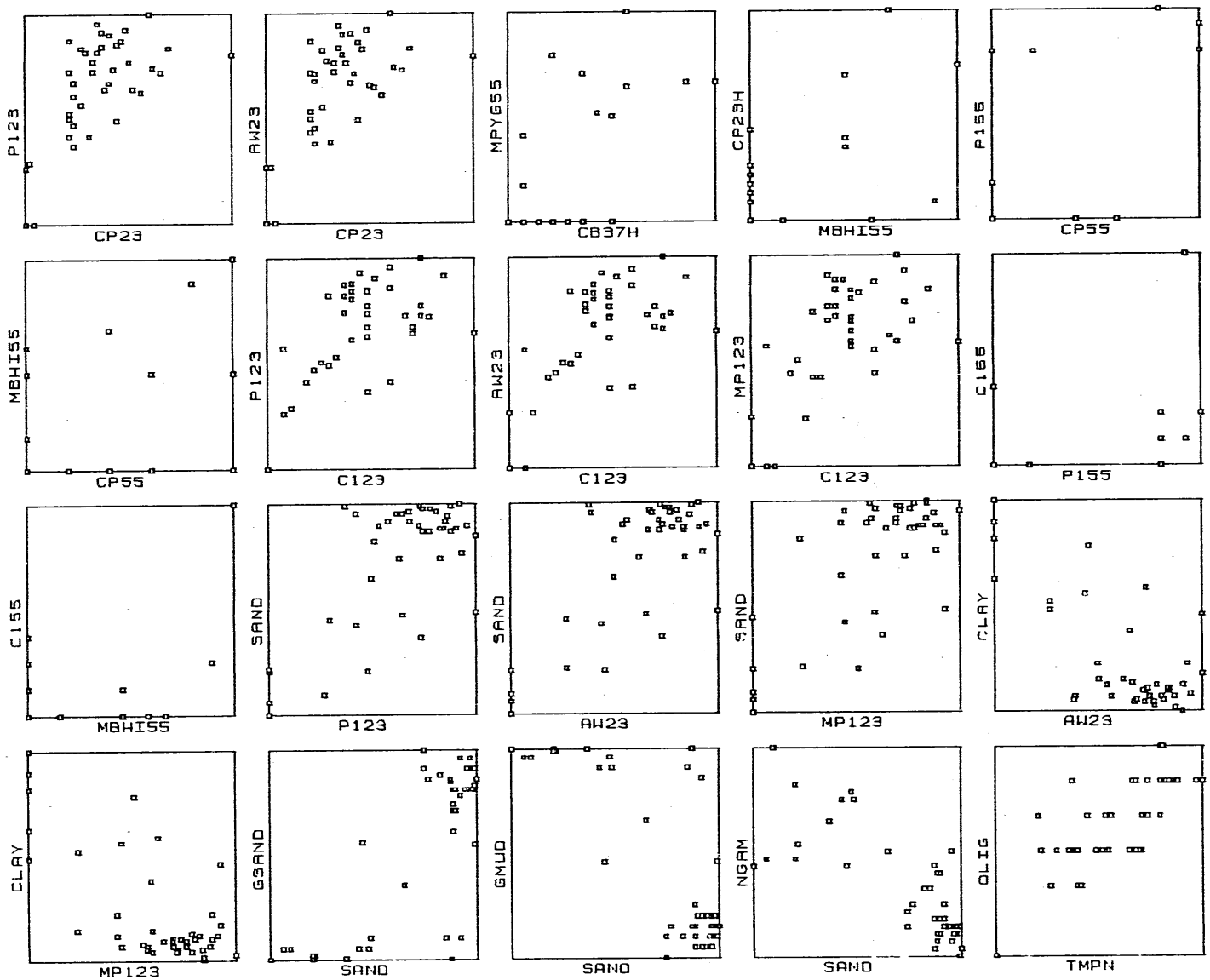
0.8 > |r| > 0.7

FIGURE 7



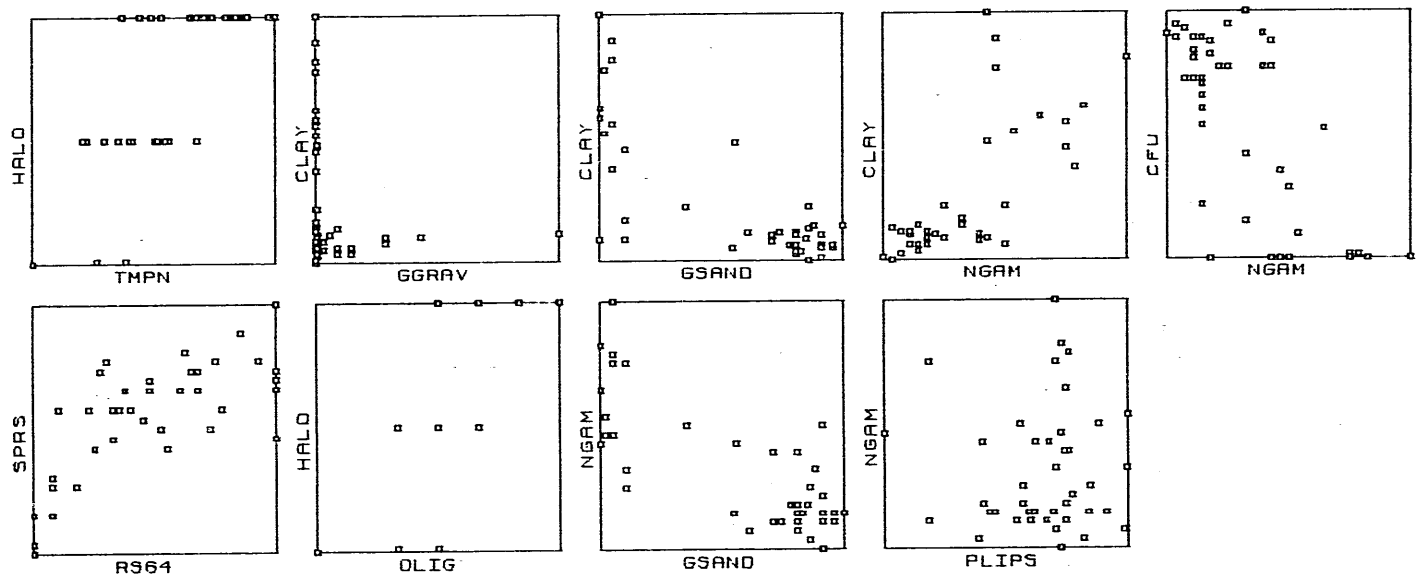
0.8 > |r| > 0.7

FIGURE 8



0.8 > |r| > 0.7

FIGURE 9



0.8 > |r| > 0.7

PRINCIPAL COMPONENT ANALYSIS

Interpretation of variable interactions through examination of only bivariate correlations remains formidable when we look at them one at a time. In our examination of the matrix of correlations we have seen that there are goups of bivariate pairs that tend to be mutually high. E.g., the upper left triangular block containing pore water chemistries, the central block containing several similar microbiological variables, etc. However, while these patterns are inherent in the data they are easily discernable only because of the way I ordered the variables to begin with. If I had randomized the variable list before preparing the matrix of correlations few patterns would have been apparent. Thus, the multivariate technique of principal component analysis introduced in the last investigator Update (March 15, 1988) provides a way of discovering the hidden patterns in the correlation matrix. The purpose is to discover natural variable associations from which one may infer biogeochemical characteristics. That is, among the one hundred variables we wish to discover the minimum number of independent factors that determine the system characteristics. This dimension reduction allows us to simplify our dicussion of the system under study. Because principal component analysis finds the "natural" associations among the variables, unbiased by our notions of what ought to be, our inferences are strengthened.

Mathematically principal component analysis begins by treating all of the variables as being equally important, regardless of the measurement units. It assumes that the variance contribution from each variable accounts for about 1% (1 X 100%/110) of the total system variance.

> Note that when we speak of variance we mean the total spread in the data. It is not a measure of precision or measurement error. The total variance observed is the sum of true differences among the samples plus contributions from measurement bias and random measurement errors. Principal component analysis allows us to separate these contributions so that we can focus attention on important differences.

After extracting the principal components the variance is redistributed, according the patterns of correlation, into new compound variables called principal components. The matrix of correlations from the present data matrix indicates that there is measurement redundancy, i.e. some of the measurements are conveying the same information and should be grouped. Indeed, our preliminary examination shows that instead of having to look at 110 variables to see 100% of the total information, we can look at about 10 principal components and see 75% of the relevant variance. Instead of having to look at 5995 scatter plots to see all of the information conveyed by the data, we only have to look at about 45 (10X9/2) principal component versus principal component plots and we will be sure of seeing 75% of the information. This constitutes a valuable reduction in the dimensionality of the problem. Below I have tabulated how the variance has been redistributed into the first few principal components.

TABLE 3

VARIANCE REDISTRIBUTION

Principal Component	Eigenvalue	% of Variance	Cum.%
1	29.0	26.4	26.4
2	14.7	13.3	39.7
3	9.6	8.7	48.4
4	5.9	5.4	53.8
5	5.1	4.7	58.4
6	4.0	3.7	62.1
7	3.6	3.3	65.4
8	3.4	3.1	68.4
9	3.0	2.7	71.2
10	2.5	2.3	73.5

Since each original variable contributed an equal amount of variance to the system, a single variable which had zero correlation to any other variable would constitute an eigenvalue of 1.0. The table shows the first principal component is equivalent to 29 of the original variables. Twenty-nine variables are measuring the same fundamental behavior. The second principal component is equivalent to about 15 variables and accounts for a different 13.3% of the total information in the data base. As indicated in the last Update, the product of the analysis yields objective quantitative numbers (called factor or principal component scores) that permit comparisons among the samples. Thus, if we were to prepare a plot of the scores (location of the 43 original samples on the new compound principal component axes) we would be displaying 39.7% of the total information on ONE plot, PC1 versus PC2. It is easy to see why we exploit this multivariate technique for summarizing data from large data bases. In the next Update we will examine a series of principal component versus principal component plots to see how the samples relate to one another and to determine whether clusters of behavior among the samples exist. In this Update we will graphically examine the spatial (vertical profile) variability among all of the samples to determine how behavior differs among the wells. These multivariate plots provide an exploratory power for viewing the totality of the data that would be too cumbersome by looking at a series of individual variable versus depth plots.

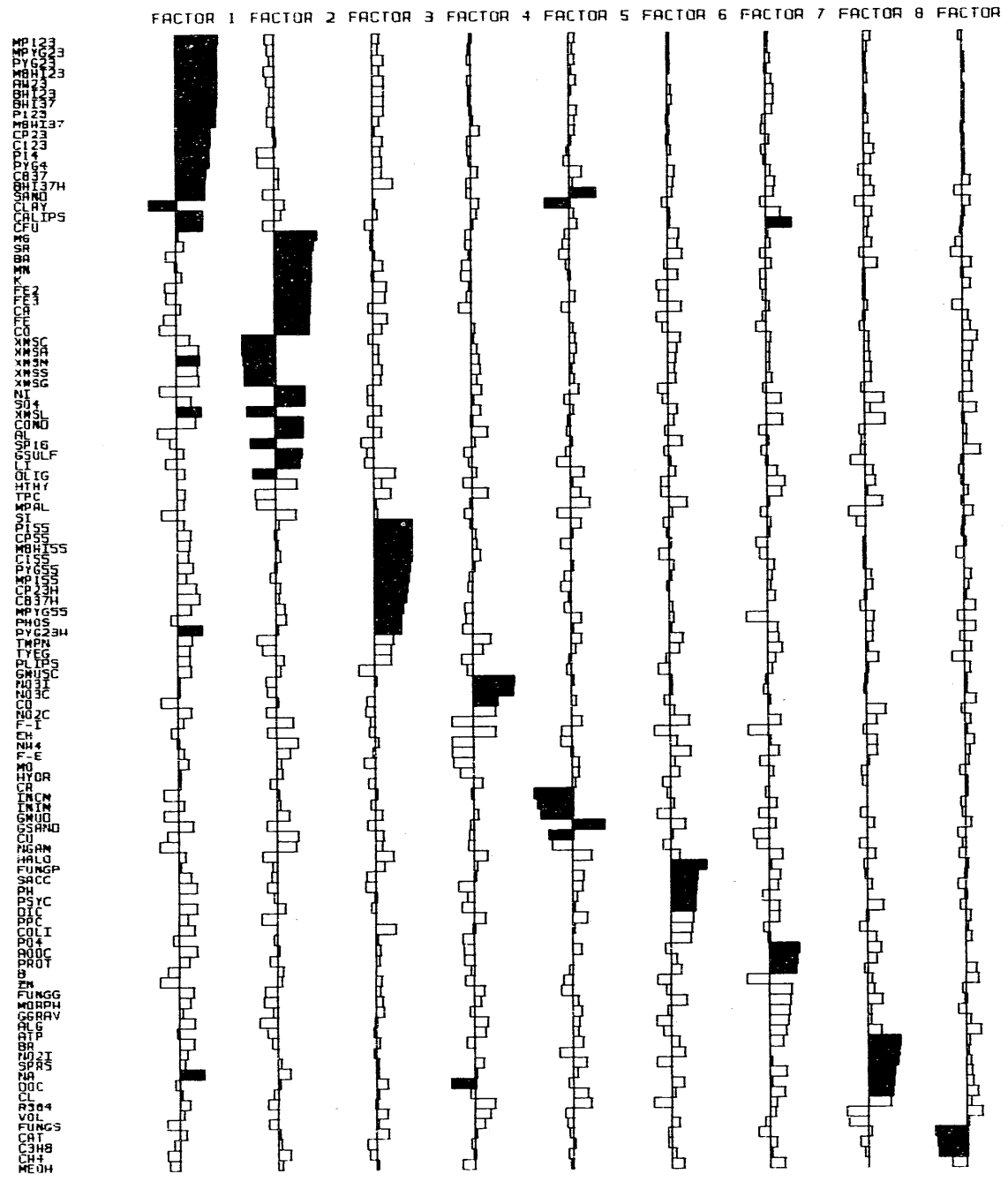
Most people who are unfamiliar with these techniques have little trouble examining multivariate plots and seeing relationships among the samples. However, they find it difficult to understand the meaning of an axis that represents 29 variables simultaneously. They are more comfortable thinking about how each individual variable relates to these complex axes. Fortunately the technique provides a way to do that. The contribution of each individual variable to the compound variable is provided by the factor or principal component loadings. The magnitude of a variable's loading on a factor indicates its relative importance compared with other variables on the same factor. In addition, numerically a variable's loading is the correlation coefficient between the variable and the compound variable. Thus, if a factor has a loading of 0.8 for variable A, and -0.4 for variable Q, it means that variable A is twice as important as variable Q. In addition, it means that variable A is highly correlated with the compound variable and one could be fairly well predicted from the other. Variable Q is inversley, and relatively weakly, correlated (r=-0.4) with this principal component. Thus, by examining the loadings of the original variables on the principal components we may explore the nature of the variable associations and begin to interpret their meaning in the context of our knowledge of their individual significances.

Figure 10 shows the relative magnitude and sign (left=neg., right=pos.) of the loadings for the first nine factors. Loadings greater than magnitude 0.5 the are shaded to focus attention on the major contributors to the factor. Note that every variable loads on every principal component. However, most variables seem to have large loadings on only one factor. The vertical order of the variables on this plot is determined by its loading on the factors. The loadings of MP123 through CFU are largest on Factor 1 and small on all other factors. Variables MG though SI are the largest loadings on Factor 2, and are lower on all other factors. Thus, loadings indicate which variables are related to one another and tend to be associated in the data base. The loadings plot summarizes all of the associations that may not be immediately obvious from visual inspection of the correlation matrix.

I will illustrate how the loadings reveal the structure of the correlation matrix by examining a single variable, CLAY. Examining the correlation matrix indicates that clay content is inversely correlated to several microbiological variables (BHI23, PYG23, P123, AW23, MPYG23, MP123), and positively correlated to three pore water chemistry variables (INIM, AL, SI). Now examine the loadings plot. Note that clay is among the largest loadings (contributions) on Factor 1, but its relationship to all of these largest loading variables is negative. Closer inspection of the correlation matrix will also show that all of the variables with large positive loadings on this factor are positively correlated to each other on the correlation matrix. If you now examine how CLAY loads on the other factors you will note

that it is not very large on any other factor. However, CLAY makes its next largest contribution on Factor 5. It is inversely correlated to SAND (As we would expect). And directly correlated to INIM, INCM, GMUD, which are also negatively loaded on this factor. Thus, we can see that individually we can examine the variable contributions on these factors. But more importantly, the groupings are of great interpretive significance.

FIGURE 10 FACTOR LOADINGS



INTERPRETING THE FACTORS

While performing these computations is relatively straight-forward, it is not expected that the investigators would be inclined to do them. Indeed, there is a great wealth of information to be gleaned from the detail of the computational output. However, this too may be beyond the interest of the investigators and I have undertaken the task of trying to call your attention to some of the significant findings uncovered by my examination of the subtlties. I must emphasize that these techniques do not answer questions, but they help to reveal significant relationships that require our attention. The relationships among samples or among variables are REAL. Some of them are trivial because they are induced by measurement artifact, but they exist and we must be aware of the artifacts. Other associations are trivial because they represent information that is obvious or which has been known for a long time. But I must emphasize that when the obvious is "discovered" by this purely mathematical artificial intelligence, it is good evidence that the technique is capable of "discovery" of any sort. Other associations that I may call to your attention are equally real, but their significance escapes me because I have insufficient microbiological, geological, or chemical knowledge to appropriately interpret what has been "discovered".

IN ORDER FOR ANY OF THIS TO BE OF VALUE I MUST RELY UPON YOUR EXPERTISE TO INTERPRET WHAT THE DATA REVEAL.

I will procede with a factor-by-factor description of the loadings. The purpose is to illustrate the how one uses these exploratory techniques and to come up with a name for the variable associations. Again, the principal components represent the number of fundamentally different behaviors that are present in the data. It is important to identify the nature of these behaviors in the context of the science. In some cases my suggested interpretation may be off-the-wall because because of my ignorance. Please do not discount the fact that the associations exist. Instead, apply your knowledge and come up with a better interpretation and name for the factors. I have offered this little apology because I have played a similar role in other interdisciplinary projects and valuable time was lost over the issue of interpretation. The power of the methodology was incorrectly discredited because of my incorrect interpretation. These are your data and you are best equiped to interpret them.

FACTOR 1

Principal component 1 accounts for 26.4% of all the variance in the system. It is equivalent to 29 of the original variables. The simultaneous large loadings of several variables shown in the factor loadings plot (Figure 10) indicates measurement redundacy among the variables MP123 through CFU. Please note that in this context, redundancy does not mean superfluous. It means that the correlated variables may be predicted from one another, that there MAY be a causal relationship between them, or that some mechanism links them. Note that most of the heavily loaded variables are microbiological enumerations and colony diversities on different media, but that they are all from temperature treatments between 4 and 37 degrees Celsius. For the time being I will call this Microbiology 1. (I don't know what the correct terminology is, but it seems that we should incorporate the fact that all of these variables are from the moderate temperature range. Is the correct term mesophiles, as opposed to thermophiles which appear on Factor 3?) It is also significant that the only non-microbiological variable with a substantial contribution on this factor is the positive correlation with % SAND and an inverse relationship with % CLAY. That tells us that regardless of which hole you look at, high clay content (& low sand) indicate low microbe counts and diversities. Furthermore, this is probably not the effect of moisture content because the moisture variables (INIM & INCM) have low loadings on this factor. They are more heavily loaded on Factor 5. Now for some subtilties. Few pore water chemistry variables make any contribution to this factor. That may be interpreted as an indication that pore water chemistry has little to do with the numbers and kinds of organisms that appear in this sampling of the subsurface. Also note that there are two "bursts" of intermediate sized loadings that parallel

the variable loadings for Factors 2 and 3. This is a common occurence on the first principal component of many data sets. It indicates a secondary relationship between the heavily loaded variables and these other factors. In this case the substructure makes sense, we should expect that the responses as measured by the mineral salt carbon sources (XMS_'s) should bear some relationship to the other Microbiology 1 variables. In addition, the substructure implied for the Factor 3 variables (thermophile indicators) might be expected to bear a relationship to this factor.

FACTOR 2

This principal component accounts for 13.3% of the total variance in the data base. Most of the variables loading heavily on this factor are pore water chemistry variables. For this reason I will suggest that we call it the geochemical factor. This factor is similar to factors that I have seen in examining other deep subsurface waters. The factor may suggest that the water content reflects an equilibrium (or near equilibrium) between the mineralogical materials with which the water is in contact. Thus, the waters are an indirect indicator of the subsurface mineralogy. Its appearance as a separate factor indicates that there is a large variance among the samples, which means a variety of geochemical environments are represented. This is consistent with the relatively diverse subsurface lithological units. Unlike Factor 1, which appears to contain microbiological variables only, this factor contains contributions from several microbiological variables. The first group loading with the geochemical variables is the mineral salts with carbon sources (XMS_'s). These variables all load negatively on this factor. These variables are slightly loaded on Factor 1, as indicated earlier. Simultaneous loading of variables on two factors usually indicates that there are two sources of variance in the variables. In this case, it probably indicates that the counts on XMS_'s are "dependent" on two things. High XMS_ counts require large (first factor variable) counts and low geochemical counts. I can not suggest a rationale for this behavior. In addition, OLIG, TPC, and MPAL also load negatively with the chemical variables. Only one microbiological variable, tritiated thymidine utilization (HTHY) is positively correlated with this factor. The significance of this observation is not clear since this variable has no high correlations with other variables. Thus, I would hesitate to over-interpret this feature.

FACTOR 3

Principal component 3 accounts for 8.7% of the total variance in the data base. Major variables loading on this factor are predominantly enumerations and colony diversity counts from the high temperature incubations and 23 and 37 degree incubations following heat shock treatment. Therefore, I will call this the thermophile/spore former factor. The marginal positive loadings of these variables on Factor 1 shows that they are only slightly related to this microbiological factor. It is interesting to me that the principal components analysis has "discovered" the difference between these variables and the other microbiological variables and placed them on separate independent factors.

FACTOR 4

Principal component 4 accounts for 5.4% of the total variance in the data base. It consists mainly of pore water nitrate, nitrite, EH, ammonium and fluoride measurements. I will call this the nitrogen factor. Nitrate measured by colorimetric analysis (NO3C) and ion chromatography (NO3I) are loaded together, as they should be. They are redundant measures of the same thing (correlated at r=0.97). This redundancy is truly duplicative. However, they are also related to nitrite by colorimetric (NO2C). This indicates there is a chemical relationship between oxidized and reduced forms of nitrogen. Additional evidence of the redox relationship is that the redox potential (EH) also loads on this factor. Since nitrite by ion chromatography (NO2I) does not load with nitrite by colorimetry we may conclude that the its variance contribution is due to some other factor (See Factor 8). I am at a loss to explain the appearance of fluoride measures on this factor. But the absence of any

significant loadings from microbiological variables on this factor indicates that the nitrate redox relationship is geochemical in origin and not due to microbial activity. While all of the other factors appear to be representing general behavior among all samples, this factor appears only to account for the behavior in core P-29. It is the only one that seems to have substantial nitrate in the pore waters. In the last Update I suggested that this is an interesting observation that will require examination when the fourth hole is completed.

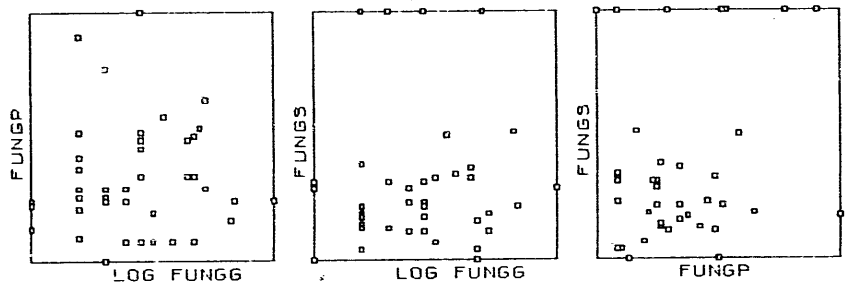
FACTOR 5

Principal component 5 accounts for 4.7% of the total variance in the data base. It is a moisture factor. As expected it has loadings from direct quantitative measurements of sand and clay and the semi-quantitative well log observations (GMUD & GSAND). Also note that the well log N-gamma measurements appear on this factor. Pore water copper (CU) is the only chemical variable loading on this factor. Its positive correlation with moisture, sand, and copper is interesting but I need some help from a geologist to interpret its significance. I would emphasize that this moisture factor appears to be unrelated to all microbiological variables. And, while sand and clay are related to moisture, sand and clay content only are also related to Factor 1 (microbe 1 factor). As indicated in dicussion of Factor 1, this may be evidence of a surface area/particle size relationship with microbial content.

FACTOR 6

Principal component 6 accounts for 3.7% of the total variance in the data base. It consists of microbial variables including fungal counts (FUNGP) Saccharolytic, Psychrophylic, and coli enumerations. Pore water PH, dissolved inorganic carbon (DIC), and phosphate also load on this factor. I need a microbiologist to interpret this factor. I also think that we need to determine why the fungal counts from Phelps, Ghiorse, and Suflita (FUNGP, FUNGG, FUNGS) do not appear on the same factor and why they are so poorly correlated (all less than 0.1, see scatter plots Figure 11). Where they determined differently, or am I assuming the wrong thing about the nature of these variables based on their "names" in the data reports that I have?

FIGURE 11



r = 0.046 r = -0.042 r = 0.188

Note that all three variables are expressed as logarithms of the enumerations. However, FUNGP and FUNGS were supplied as logs and FUNGG was supplied as enumerations and transformed here.

FACTOR 7

Principal component 7 accounts for 3.3% of the total variance in the data base. When I first looked at this factor I called it the Ghiorse factor because five of the ten strongest loadings are Ghiorse variables; AODC, protozoa (PROT), fungi cfu (FUNGG), algae (ALG), and ATP. Two pore water chemistry variables load on this factor, boron being positively correlated and zinc negatively with the microbiological measures. I don't know how to interpret this factor, but I beleive that it is important because the microbiological variables on this factor are not "redundant" with the other microbiological measures. That means that they are providing unique information. Please give some thought to an interpretation of these loadings. As we shall see in the vertical profiles this behavior seems to be uniquely associated with the upper strata of P-24.

FACTOR 8

Principal component 8 accounts for 3.1% of the variance in the data base. As I indicated earlier, sometimes a factor has no mechanistic significance in terms of the science, but that it represents measurement artifact. That is how I interpret this factor. I suspect that it represents a measurement bias on pore water nitrite by ion chromatography (NO2I) when high bromine and dissolved organic carbon (BR & DOC) are present in the samples. My own experience with this measurement is consistent with this observation, since the bromine nearly coelutes with NO2 under some conditions, and weak organic acids (DOC) can also adversely affect the quantitation of nitrite. I have discussed this with Tom Garland and I believe that he concurs. This is also consistent with the observation that the loadings for these variables is inversely proportional to the pore water volume recovered. Small amounts of water extracted mean more concentrated DOC, B,r etc. As indicated in my description of Factor 4, which represented true nitrate/nitrite characteristics, the present factor is indicative of measurement artifact and not a true measure of nitrite. It was for this reason that I suggested that we might consider dropping nitrite by ion chromatography from further consideration in the next hole.

FACTOR 9

Principal component 9 accounts for 2.7% of the total variance in the data base. It consists of Catechol and gas utilization (CH4 & C3H8) variables. For lack of a better term I call this the gas factor. I need some help in interpreting this factor as well. In some respects this factor is like the nitrogen factor (Factor 4) which identified core P-29 as unique in its nitrogen characteristics. This factor suggest the unique behavior of the two Ellenton samples in P-24. (See profile plot Figure 20).

REMAINING FACTORS

The first 9 factors account for 71.2% of the total variance in the data base. They represent the majority of the interpretable information. The remaining 28.8% of the variance is distributed among the remaining 101 principal components. For the most-part, they represent purely random associations (accumulated measurement errors) or unique factors that contain loadings from individual variables that are unrelated to all other measures. This was the goal of performing the principal components analysis, to separate the information from the noise. Thus, we have determined that the data base contains about 9 independent sources of variability among the samples. We have attempted to explain the natural associations of variables that constitute the factors in terms of their geochemical and microbiological compositions.

While we have accounted for about 70% of the total variance with the nine compound axes (principal components) you may be wondering how well we have explained the variances contributed by the individual variables. Fortunately the principal components technique

allows us to determine that. Table 4 shows the **communalities**, the fraction of each variable's variance "explained" by the chosen nine principal components. Note that in most cases more than 75% of the variable's variance content has been accounted for. That means we have adequately explained the variance of the majority of the variables. However, a few variables have been poorly (less than 50%) accounted for on the first nine principal components. These variables are loaded on the 10th through 110th principal components and have not been included in our interpretation. As indicated earlier, this usually means that the unexplained variables are uncorrelated to any other variables. This could be the result of unique behavior, i.e. the variables are measuring behavior that other variables can not provide, or that the variables just measure "noise". Determining which of these judgements best characterizes the low communality variables requires additional study. Until we add additional variables and progress beyond this first-pass look at the data base I would hestitate to interpret the low communality variables.

TABLE 4
FRACTION OF VARIANCE EXPLAINED BY NINE FACTOR MODEL

VARIABLE	COMMUNALITY	VARIABLE	COMMUNALITY
VOL	.58536	MPYG55	.63677
INIM	.68392	MP123	.93404
INCM	.86318	MP155	.68949
COND	.88467	XMSC	.85429
PH	.80949	XMSA	.91513
EH	.70882	XMSS	.83531
DOC	.76197	XMSL	.85949
DIC	.74415	XMSG	.86519
NH4	.62720	XMSM	.88408
NO3C	.81759	CB37	.59334
NO3I	.86786	CB37H	.77678
NO2C	.64369	CP23	.62142
NO2I	.65523	CP23H	.81188
F#E	.65403	CP55	.76895
F#I	.76585	C123	.61588
CL	.57325	C155	.87063
BR	.80657	SAND	.88108
PO4	.33958 poor	CLAY	.86484
SO4	.86984	CFU	.75837
FE3	.79756	AODC	.76166
FE2	.88707	PROT	.54995
AL	.75763	ALG	.69338
B	.51104	FUNGG	.57327
BA	.91546	ATP	.51179
CA	.87723	TMPN	.85010
CD	.54361	TPC	.67171
CO	.86915	FUNGP	.68230
CR	.13946 poor	MORPH	.65858
CU	.75675	PPC	.63131
FE	.83319	TYEG	.52834
K	.75770	SACC	.58680
LI	.57622	PSYC	.34389 poor
MG	.93474	OLIG	.69131
MN	.85982	MEOH	.44852 poor
MO	.39804 poor	CH4	.58622
NA	.81865	C3H8	.55917
NI	.76926	HYDR	.20996 poor
SI	.64313	PHOS	.75145
SR	.90750	MPAL	.67496
ZN	.71960	CAT	.67595
BHI23	.90931	HALO	.73059
BHI37	.90355	COLI	.57587
BHI37H	.71826	FUNGS	.41764 poor
PYG4	.74742	PLIPS	.47356 poor
PYG23	.91853	CALIPS	.52553
PYG23H	.77365	HTHY	.53388
PYG55	.76298	GGRAV	.33713 poor
P14	.76593	GSAND	.84119
P123	.91023	GMUD	.85683
P155	.81512	GMUSC	.32762 poor
AW23	.94059	GSULF	.58442

MBHI23	.88610	NGAM	.74898
MBHI37	.84374	SP16	.52499
MBHI55	.86159	RS64	.88332
MPYG23	.86543	SPRS	.67062

VERTICAL PROFILE PLOTS

Having determined the nine composite axes for viewing the sample variabilities we shall now examine the vertical and spatial variability by plotting the principal component scores as a function of depth in the cores. The purpose for performing the principal component analysis is to discover natural associations from which one may infer biogeochemical characteristics. But more importantly, the product of the analysis yields objective quantitative numbers (called factor or principal component scores) that permit comparisons among the samples. And this capacity allows us to graphically examine the spatial (vertical profile) variability among all of the samples. The principal component scores indicate how each sample compares to all others. Thus, if a sample has a score of -2.5 on principal component 1 we know that it is 2.5 standard deviations below the mean in the combined content of the variables that load heavily on that factor (Microbiologicals 1). If the same sample has a score of +1.0 on principal component 2 we know that it is one standard deviation above the mean of all other samples in terms of its geochemical variables. Thus, I have prepared plots of the nine principal components (Figures 12-20) described above as a function of depth for the core holes.

Figure 12 depicts the scores for Factor 1 versus depth. Scores for each core are depicted on separate axes; with negative scores plotted to the left and positive scores to the right. The points are connected with line segments between successive points (sample locations) and shaded in order to enhance recognition of vertical patterns. For example, the first sample from P-24 (depth 113) is -0.8 standard deviations below the mean of all samples in Factor 1 (Microbiologicals 1). Sample number 9 at 657 feet has the largest positive score of 1.5 standard deviations above all others in Factor 1.

If you quickly page through the nine factor profile plots you will observe that there are some gross generalizations to be made.

> One note of caution when examining these plots is that when the scores traverse between positive and negative scores they produce a node which makes it look like there is a sharp demarcation between zones. This is an artifact of the density of sampling points. Thus, the exact vertical location of the transition points may not be adequately determined from the plots. However, the general character above and below a node indicates that somewhere between successive large positive and negative scores there is a major change in the character of the variables represented by the factor.

Nearly all of the factors go through a transition somewhere between the Ellenton and Williamsburgh where there is a discontinuity. Above this transition region (Tobacco Road through Congaree) the pattern is different from the pattern that characterizes the Pee Dee through the Middendorf. This generalization is strongest for core P-24. While many of the factor scores appear to oscillate between extremely large positive and large negative, these cross-overs occurs at a geological interface. Among the three cores several regions of homogeneity (uniformly above average or below average) which extend over several geological units are apparent. These regions are summarized in Table 5.

TABLE 5

FACTOR SUMMARY OF REGIONS

FACTOR	FACTOR NAME	SIGN	CORE	LITHOLOGIC UNITS
1	(Microbilogicals 1)	(+)	P-24	Dry Branch through Williamsburgh
2	(Geochemical)	(+)	P-24	McBean through Ellenton
		(+)	P-24	Black Creek through Middendorf
3	(Theromophiles)	(-)	P-24	Williamsburgh through Middendorf
		(-)	P-29	Tobacco Rd. through Middendorf
4	(Nitrogen)	(-)	P-24	Congaree through Middendorf
		(-)	P-28	Tobacco Rd. through Middendorf
		(+)	P-29	Tobacco Rd. through Middendorf
6	(Fungi ?)	(+)	P-24	Tobacco Rd. through Ellenton
		(-)	P-24	Pee Dee through Middendorf
		(-)	P-28	Williamsburgh through Middendorf
		(-)	P-29	Pee Dee through Middendorf
7	(AODC)	(-)	P-24	Pee Dee through Middendorf
		(+)	P-29	Congaree through Black Creek
8	(Measurement Artifact)			
9	(Gas utiliz.)	(+)	P-24	Ellenton anomaly

SIGN indicates that the region is consistently above average (+) or below (-) average.
Note that because the largest variable loadings on factor 9 are negative, above average is to the left on this factor.

VERTICAL PROFILES OF PRINCIPAL COMPONENT SCORES

VERTICAL PROFILE PLOTS (Alternate Method)

The profile plots just described are useful for examining the patterns that seem to be related to vertical lithologic units. This is enhanced by simultaneously depicting the lithologic unit connections between cores and by connecting scores between vertical sampling points. An alternative method of examining the vertical profiles is to eliminate the strata identifiers and not to connect the scores along the vertical. Figures 21 through 29 depict the same factor scores in a pseudo-three dimensional pictograph. The vertical axis depicts the sample's elevation and the factor scores are depicted by the radius of a circle drawn at the sample's depth. The circles appear as ovals due to the view point perspective (-75 degrees from the vertical). Filled circles represent positive scores (above average) and unfilled circles represent negative scores (below average). This method of depicting the factor scores tends to highlight the magnitudes and signs of the scores. Thus, they emphasize some of the generalizations described above. I tend to prefer the other method of depiction, but I have included these pictograms for those who might find an additional summary useful.

HOW TO USE THIS INFORMATION

I believe that the information that I have provided will help to interpret the data that have been acquired so far. However, I also recognize that I am providing this information in a format that may be new to some of you and that it will require some effort on your part to assimilate it. Several steps will be required to exploit the interpretive power of the statistical techniques that I have initiated with this Update. Begin by examining the Table 2 to familiarize yourself with the nature of the variables included in the data base. Next, focus on the variables that you have measured (your target variables), and examine the matrix of correlations (Figure 1) to see whether the pairs of variables that you had expected to be related are indeed correlated. Check to see if any of them are correlated to variables that were measured by one of your colleagues. Are the correlations easily rationalized? Next, examine the loadings plots (Figure 10) to see which factors contain these variables. If you find that expected correlations do not appear in the correlations matrix, the loadings plot will assist you in identifying where those variables associate with others. Try to "explain" why the variables with large loadings appear with other large loading variables on a factor. Pay particular attention to the sign of the loadings relative to its mates; negative with positive means that there is an inverse relationship between the two variables, negative with negative indicates a positive correlation. Next examine the communalities (Table 4) for your target variables. The communalities show what fraction of the target variable's variance is is accounted for in the nine factor data summary. If the communalities are low (less than 60%) it is likely that the variable is measuring a unique system characteristic, or it is just too noisey to be correlated with anything. You must decide which explanation is most consistent with your knowledge of the system. Re-examine these summaries to determine whether there is evidence that would support or refute your original hypotheses.

After examining your target variables, the next step is to begin to place them in the context of the whole data base. Examine the factors that describe the whole data base. These factors are called abstract factors because they represent strictly only mathematical associations. They become interpretationally significant only when they can be rationalized in the context of the microbiology, chemistry, geology that is suggested by the variable loadings. The variables grouped together by this mathematical procedure appear together because they have proportional similarity. That is, among the 43 samples, the variables with mutually large variable loadings tend to track one another. In some cases the correlation is to be expected because the variables simply measure the same system characteristic. In other cases, correlated variables suggest that there MAY be a causal relation between them. This implies that a mechanistic explanation should be considered. Try to come up with a "name" for the factors that reflects your understanding of the variable goupings. The names assigned to the nine or ten factors replace the need to identify by name the 110 variables with which we began. Naming the factors formalizes our finding that there are only nine or ten fundamen-

tally different "things" going on in the system and helps to focus our attention on the "differences that make a difference".

Next examine the vertical profile plots to gain some insight about how these factors vary vertically within the cores. Focus on regions that seem to exhibit vertical zones of homogeniety (several succesive samples with similarly large or small factor scores). Do these regions correspond to identifiable lithological or hydrologic units? Do they generalize amomg the cores? In cases where factor scores between cores do not parallel one another, should these be interpreted as anomalies, or as indications of spatial variability? Do they suggest hypotheses to be tested in the fourth hole? Many of the patterns in the vertical profile plots are subtle because of sample spacing. Thus, discovering patterns in the vertical profiles may require viewing them over an extended time period.

Lastly, exploratory data analysis is most beneficial when the investigators and data analysts collaborate on interpretation. Therefore, I suggest that you contact me with questions and suggestions after you have examined the materials provided. If you would like additional plots or numerical details not present in this summary, please contact me.

DATA NEEDS

It is important that the data base be continually supplemented with new data as they are acquired. I am quite certain that I do not have all of the data that are available. Please contact me about any data that you have, but that does not appear in Table 2. Some of the present interpretational ambiguitites could be cleared up by data that you have.

FIGURE 12

SAVANNAH RIVER EXPLORATORY DEEP PROBE

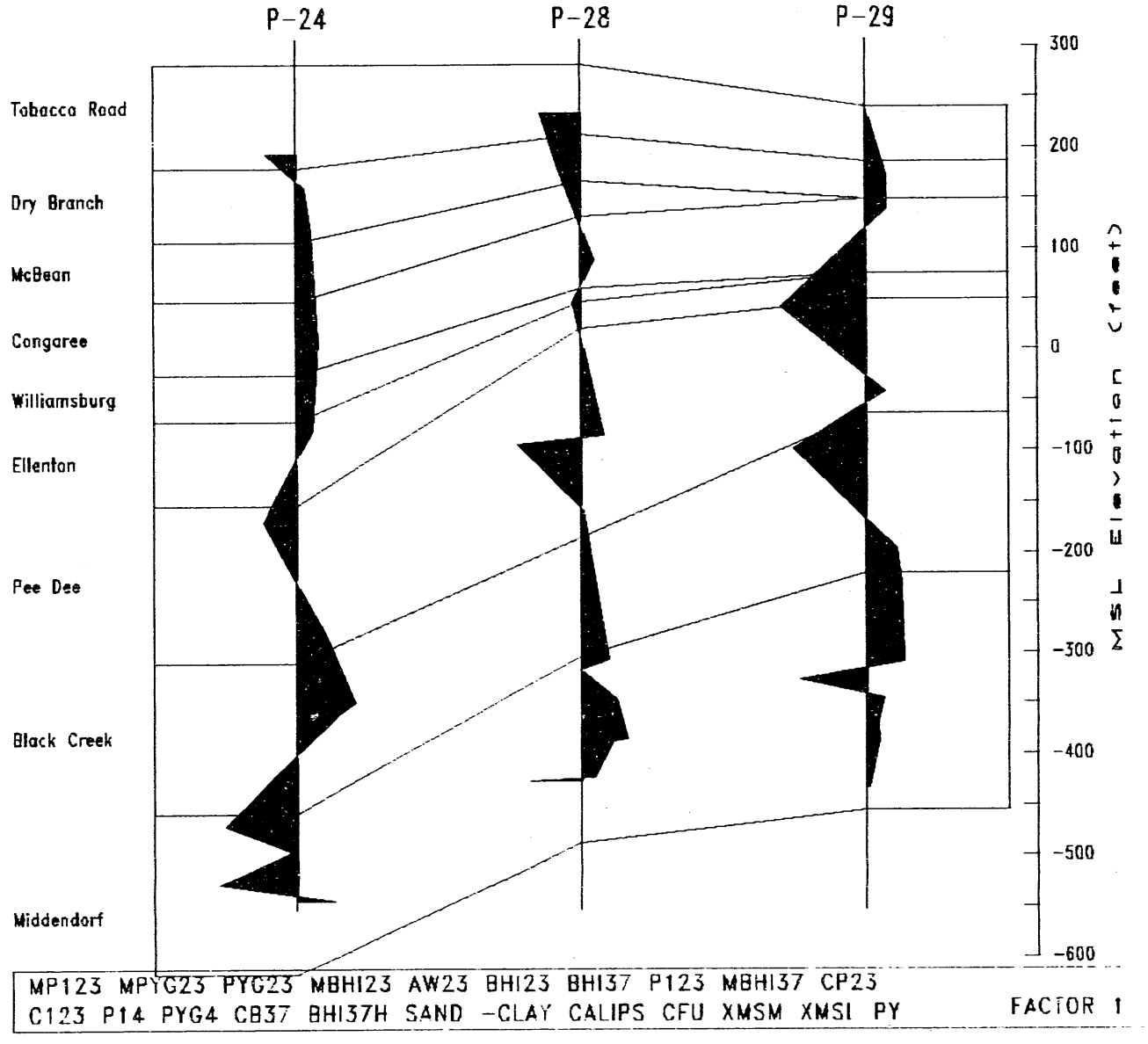


FIGURE 13

SAVANNAH RIVER EXPLORATORY DEEP PROBE

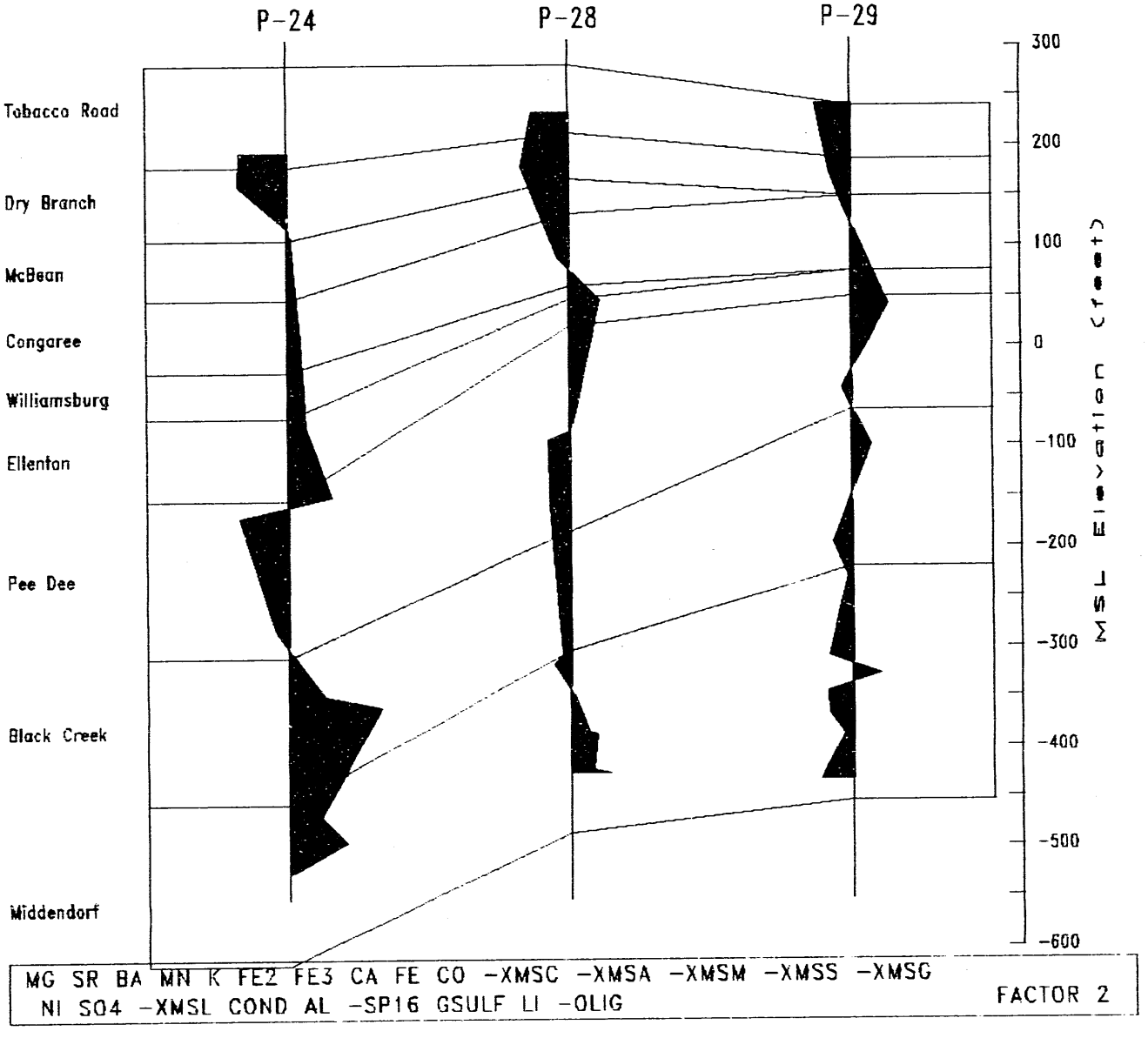


FIGURE 14

SAVANNAH RIVER EXPLORATORY DEEP PROBE

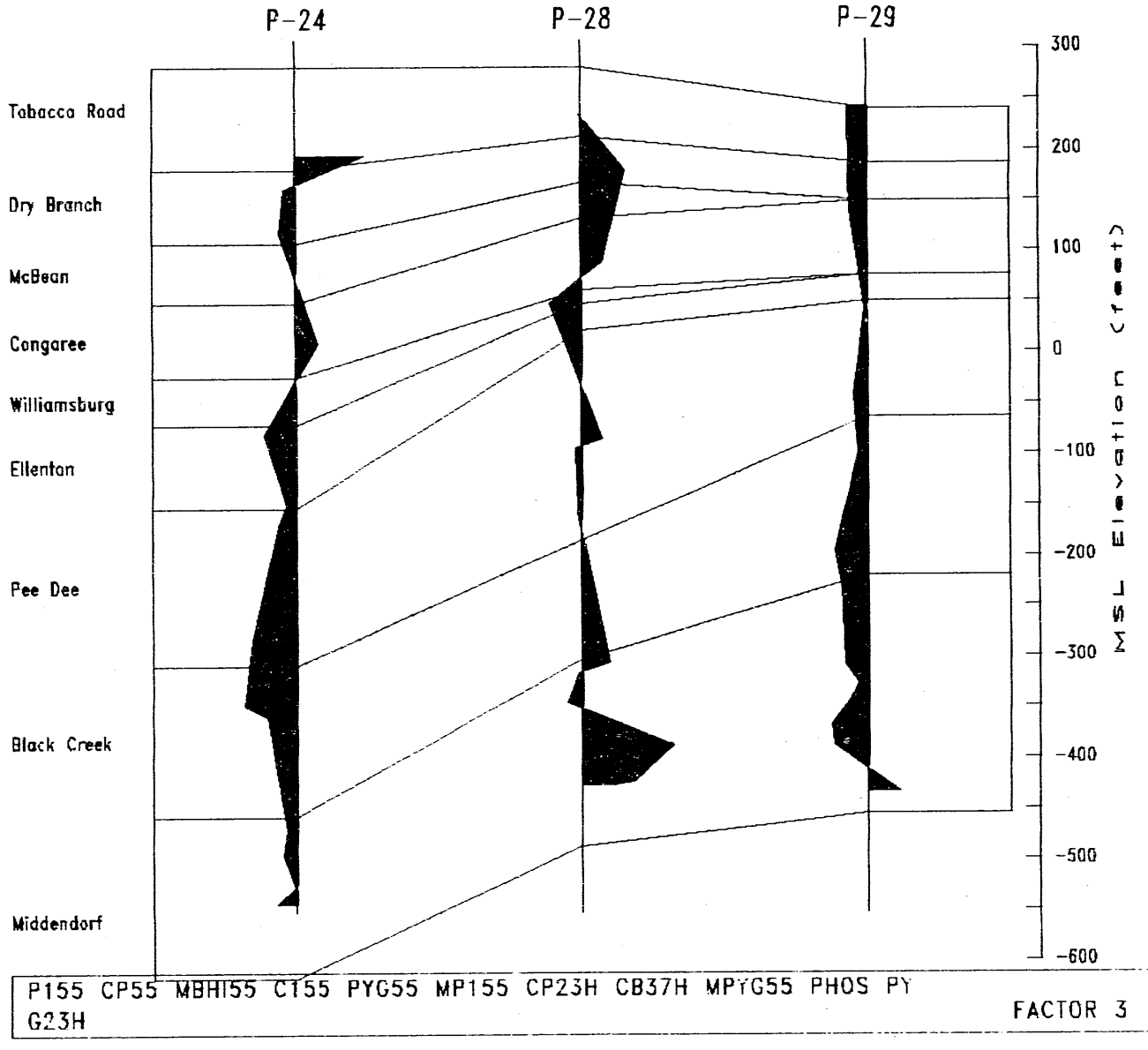


FIGURE 15

SAVANNAH RIVER EXPLORATORY DEEP PROBE

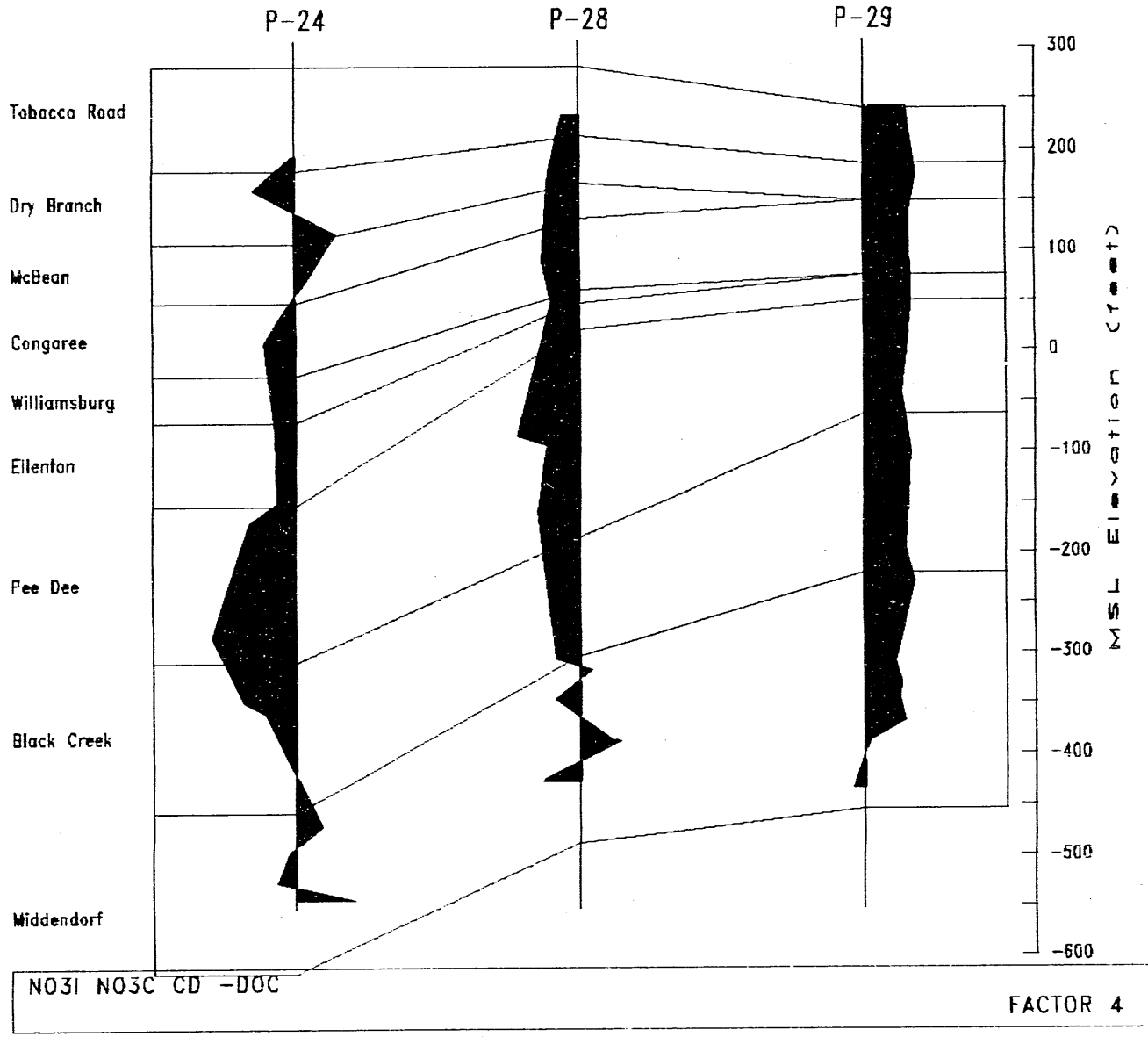


FIGURE 16

SAVANNAH RIVER EXPLORATORY DEEP PROBE

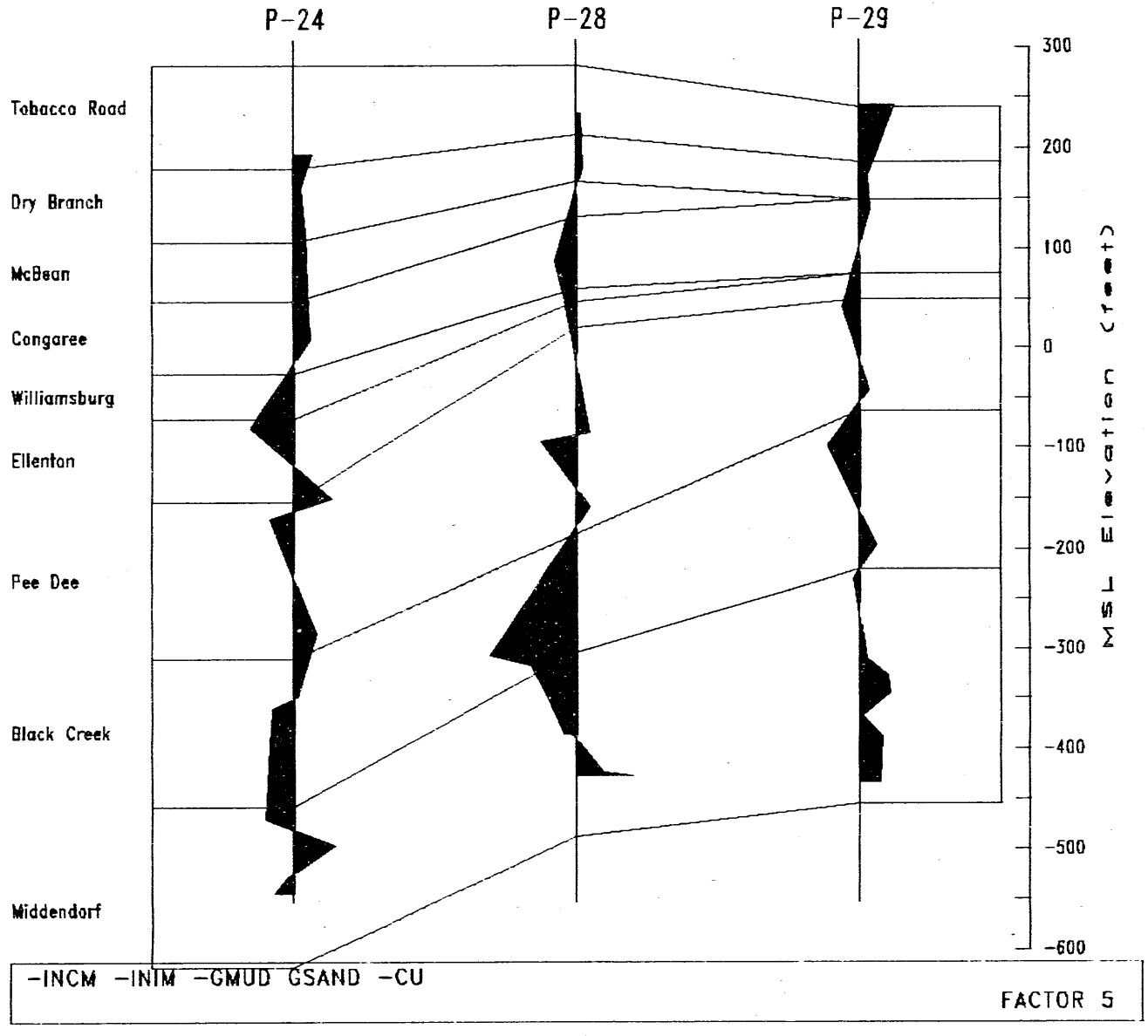


FIGURE 17

SAVANNAH RIVER EXPLORATORY DEEP PROBE

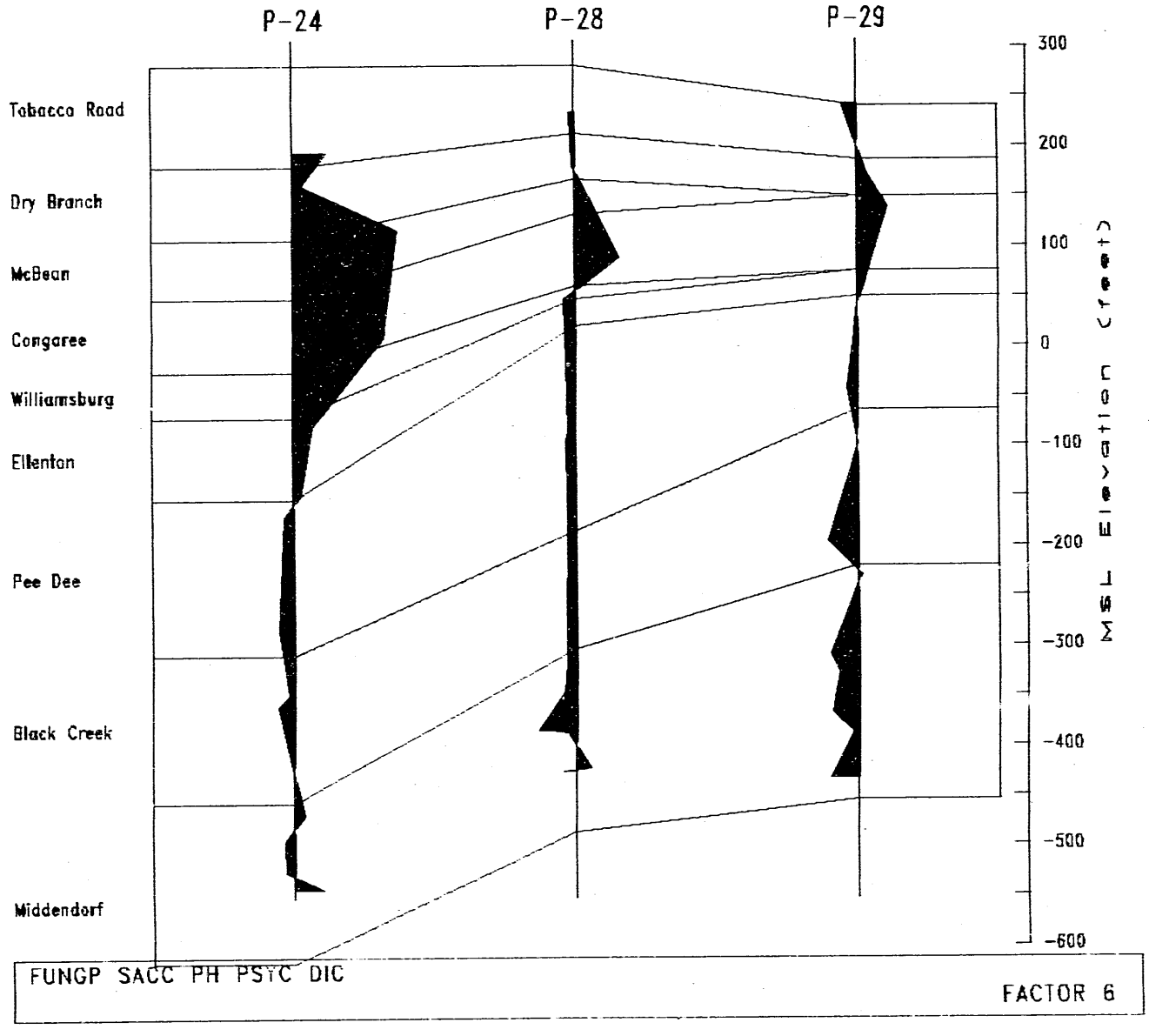


FIGURE 18

SAVANNAH RIVER EXPLORATORY DEEP PROBE

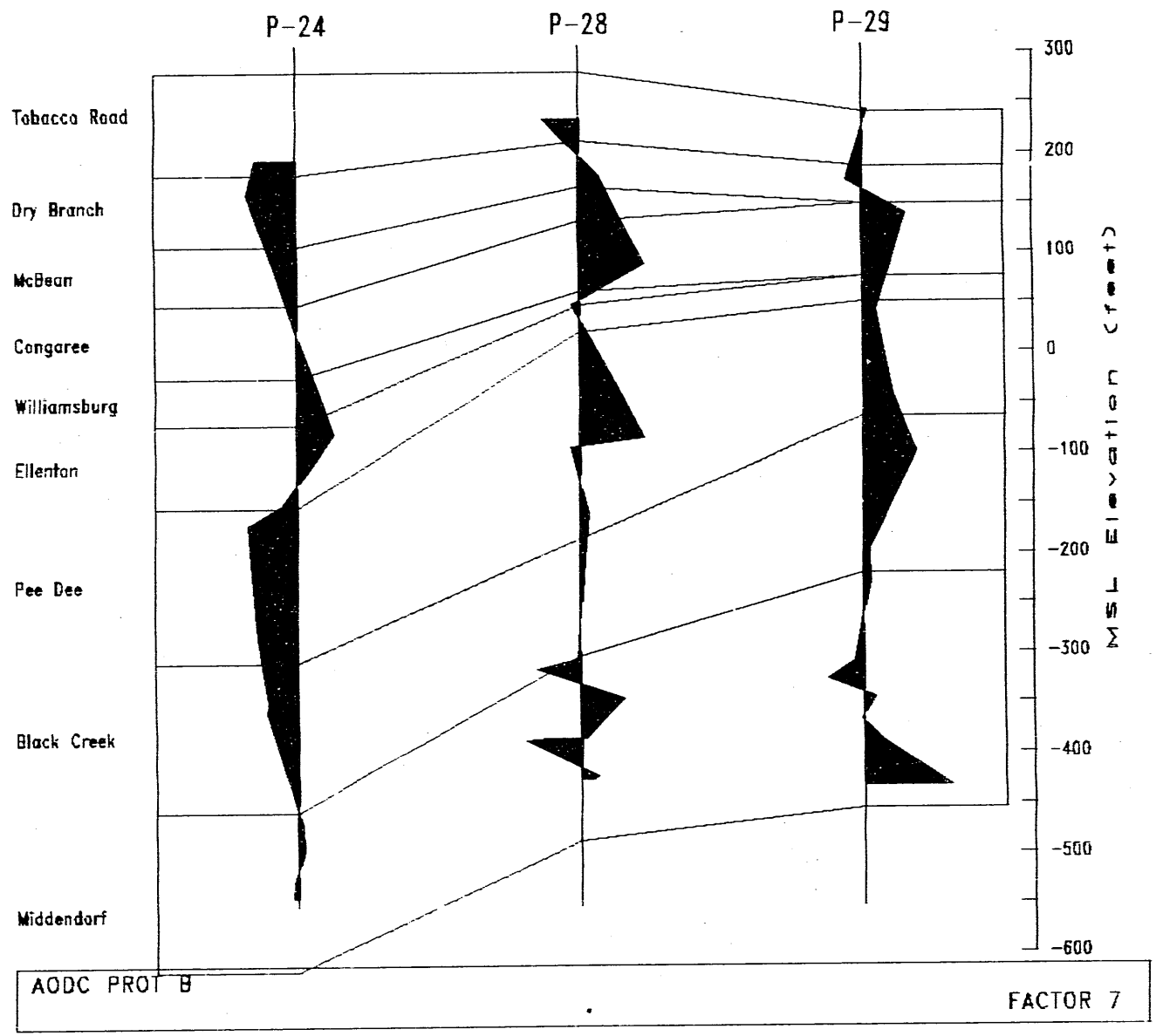


FIGURE 19

SAVANNAH RIVER EXPLORATORY DEEP PROBE

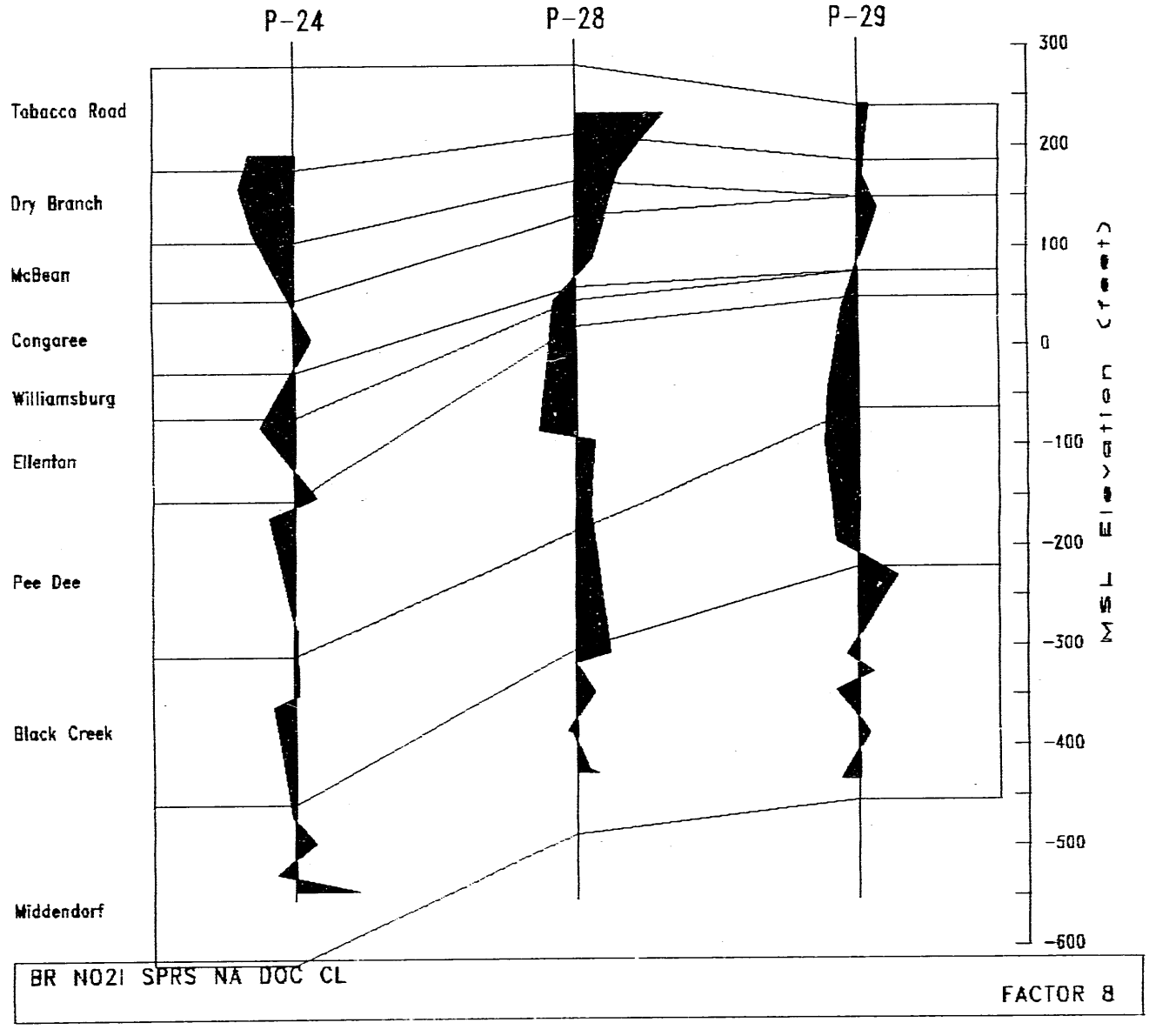
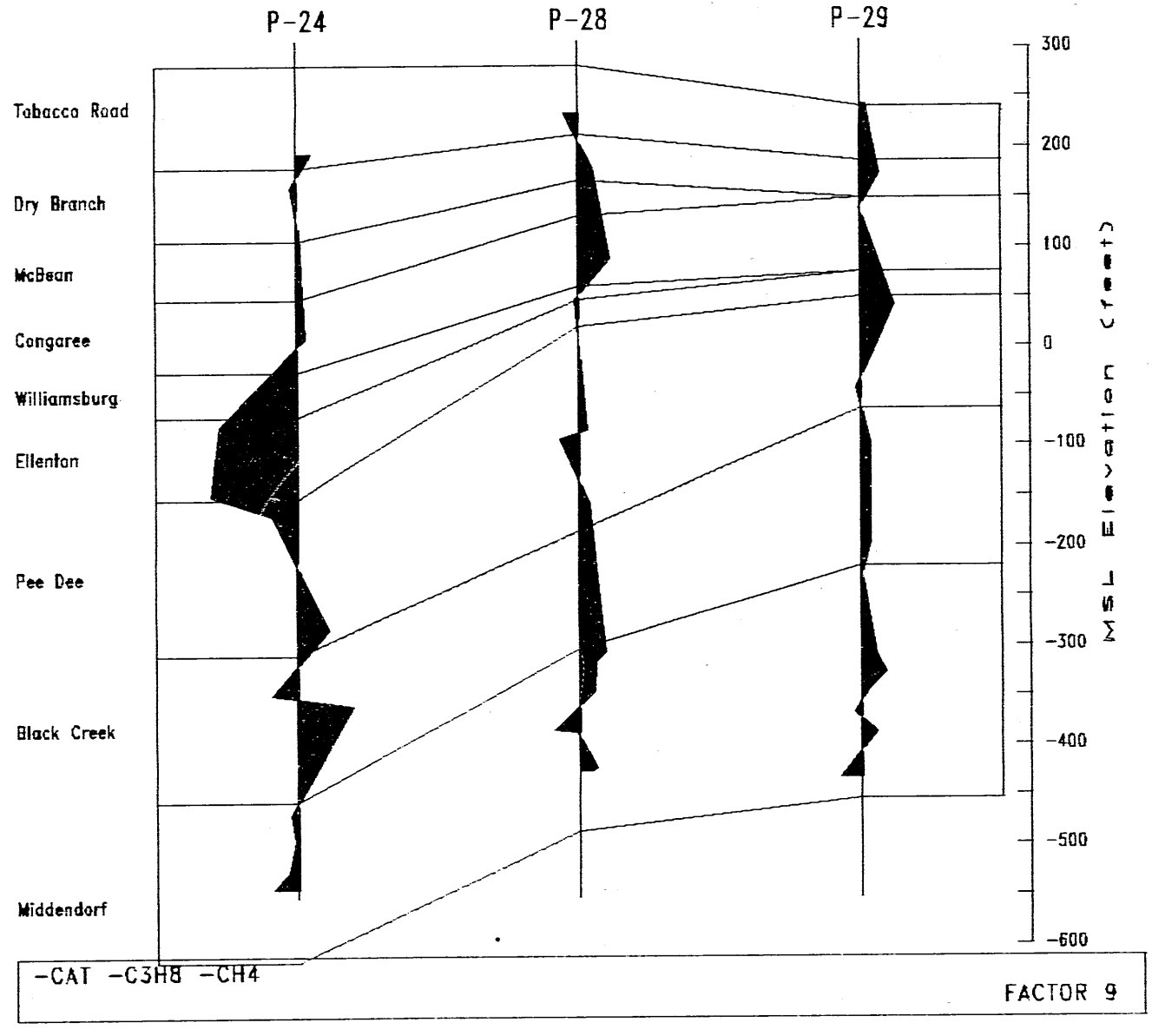


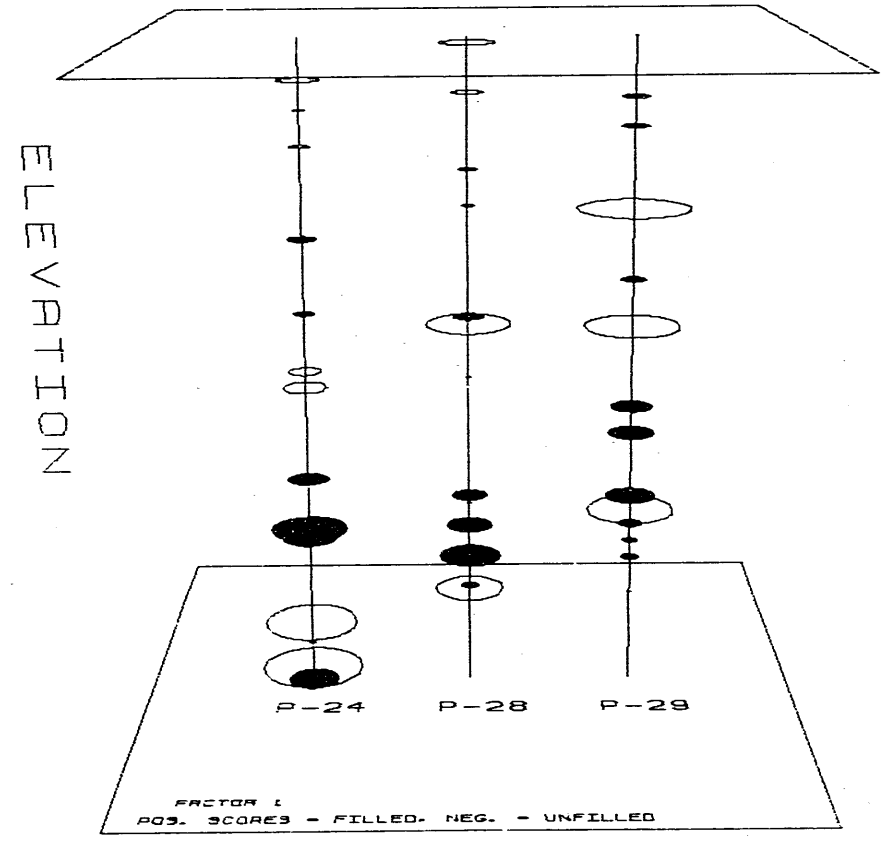
FIGURE 20

SAVANNAH RIVER EXPLORATORY DEEP PROBE



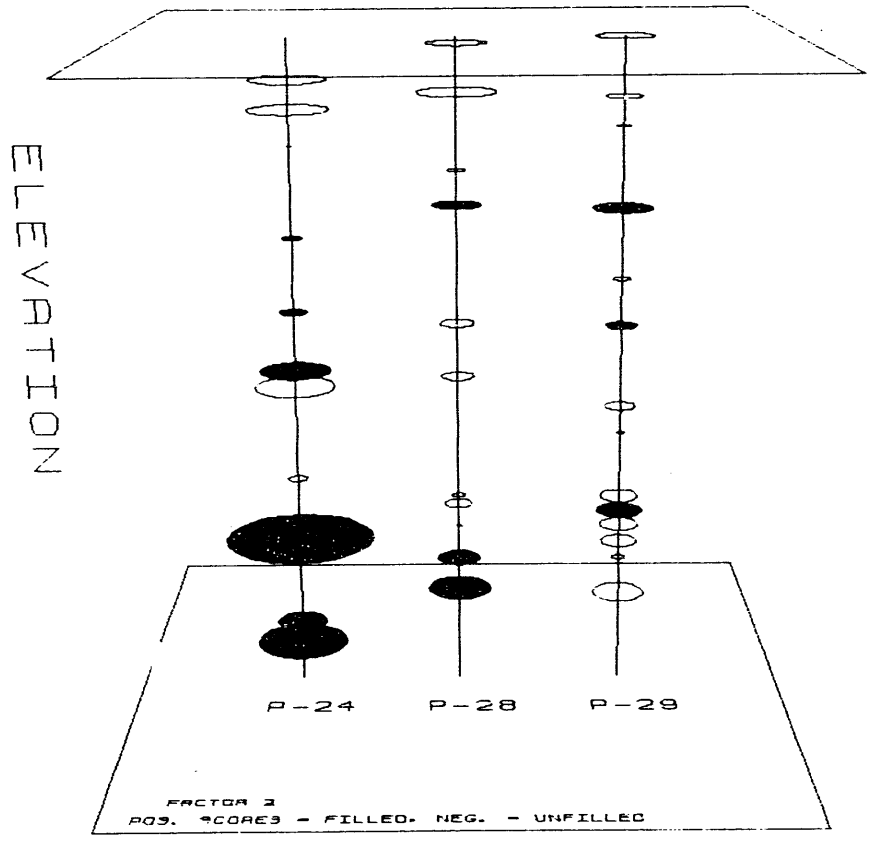
VERTICAL PROFILES OF PRINCIPAL COMPONENT SCORES
(Alternate Display Method)

FIGURE 21



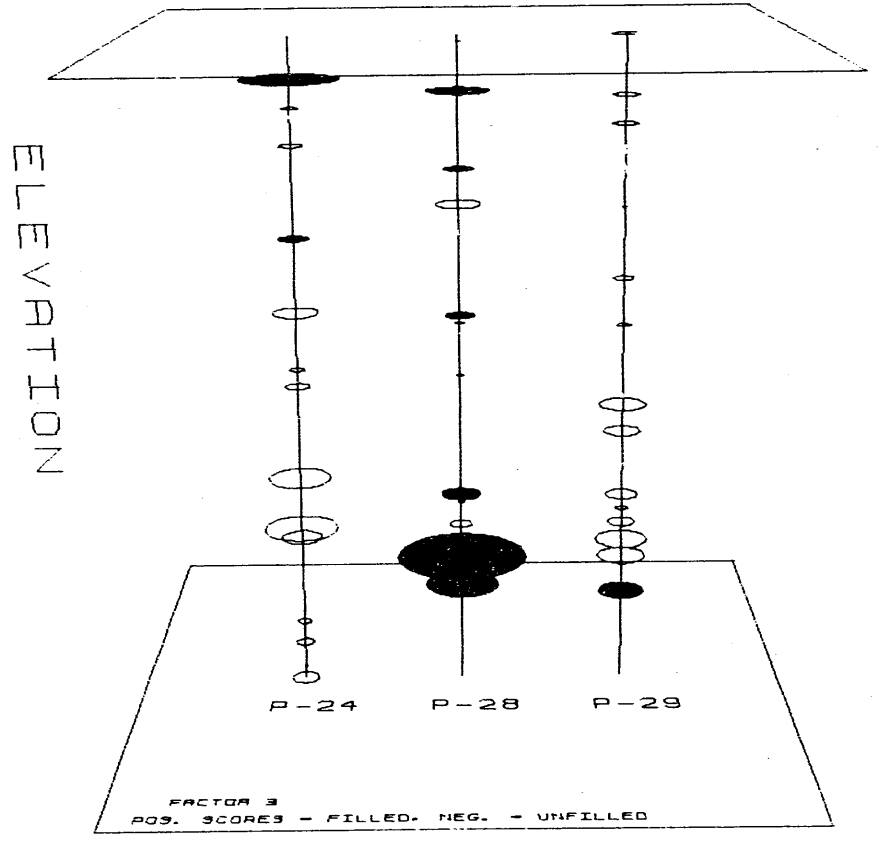
MP123 MPYG23 PYG23 MBHI23 AW23 BHI23 BHI37 P123 MBHI37 CP23
C123 P14 PYG4 CB37 BHI37H 3AND -CLA1 CALIP3 CFU XM3M XM3L P1

FIGURE 22



MG 9A BA MN K FE2 FE3 CA FE CO -XM9C -XM9A -XM9M -XM93 -XM9G
NI 904 -XM9L COND AL -9P18 G9ULF LI -CLIG

FIGURE 23



P155 CP55 M8H155 C155 P1G55 MP155 CP23H CB37H MP1G55 PHC3 P1 G23H

FIGURE 24

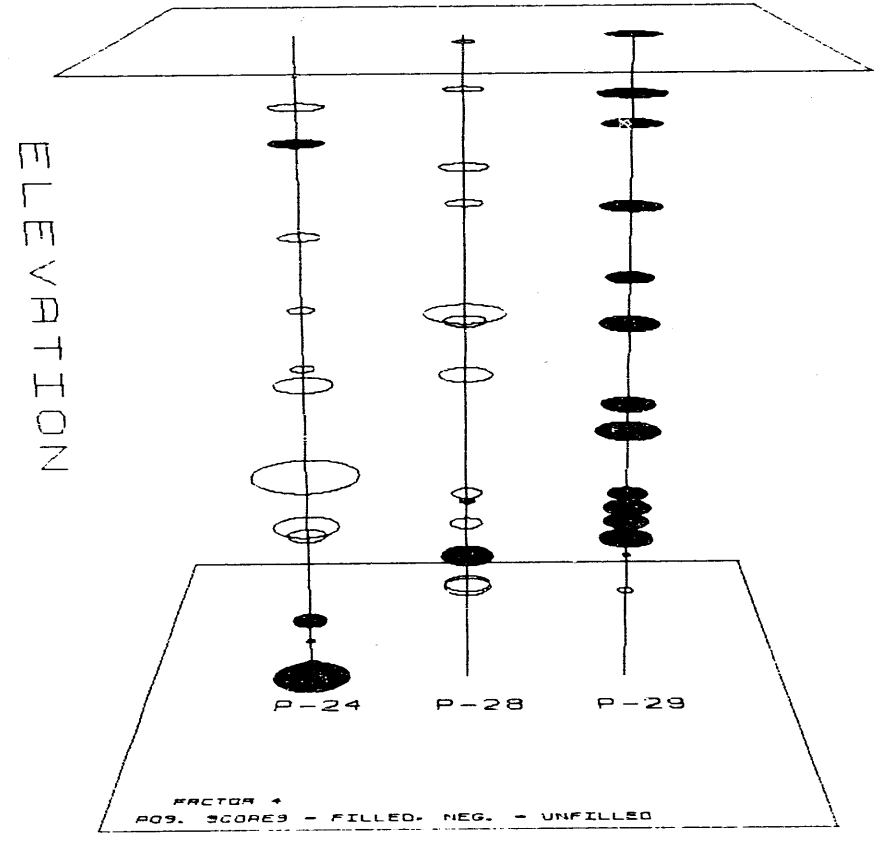
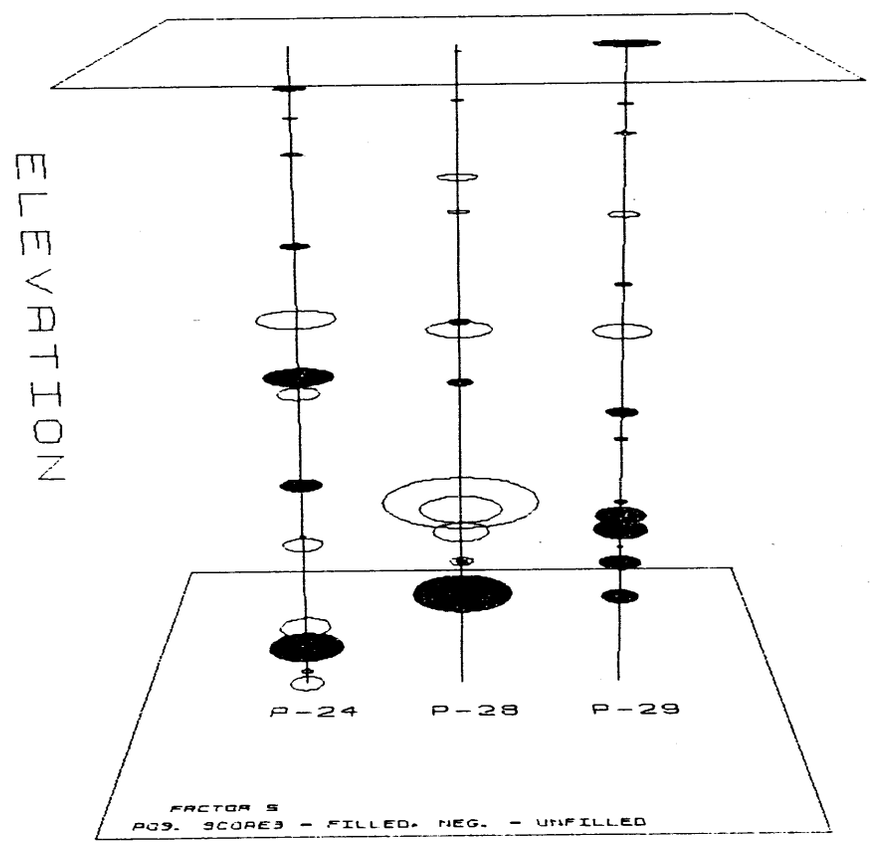
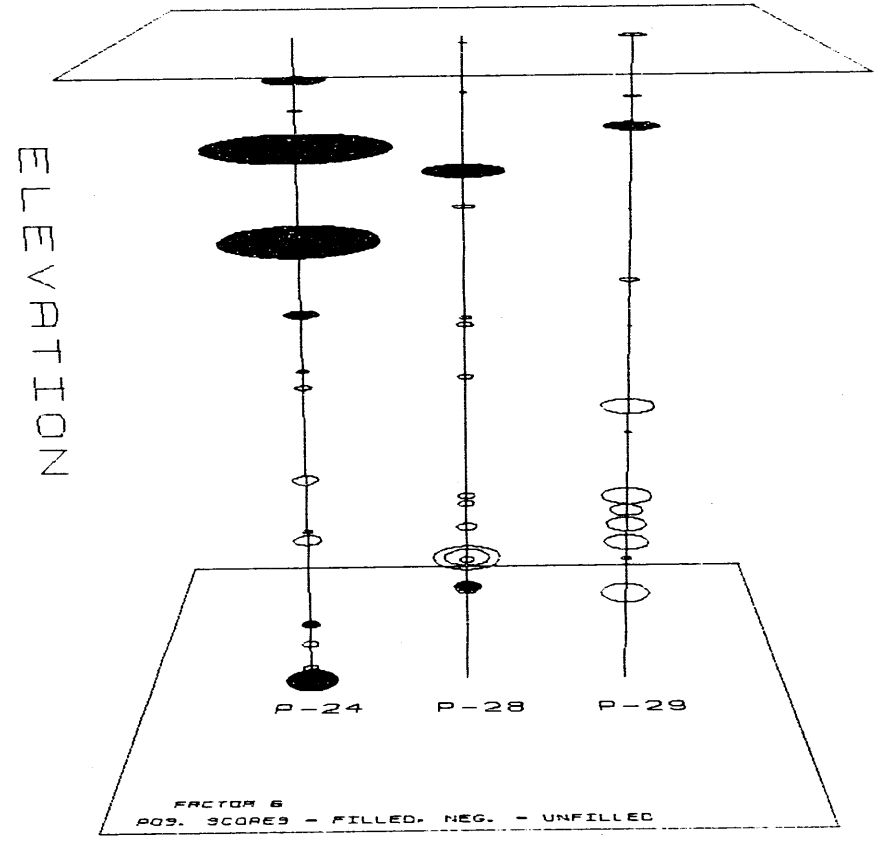

ELEVATION
P-24
P-28
P-29
FACTOR 4
POS. SCORES - FILLED. NEG. - UNFILLED

FIGURE 25



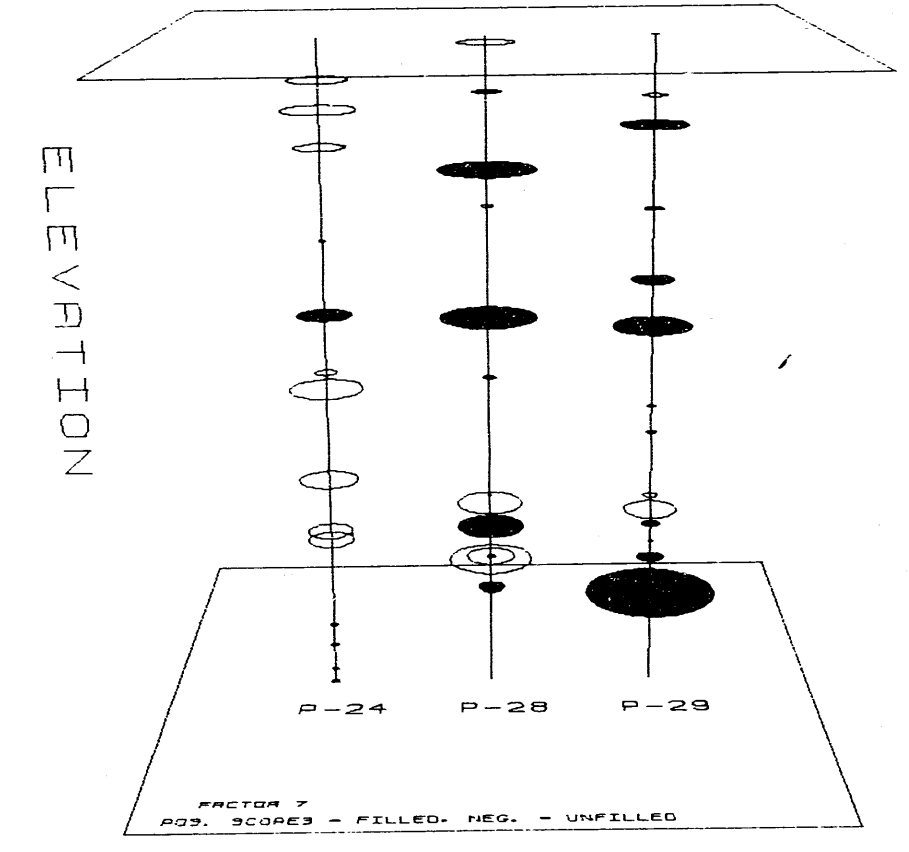
-INCM -IIIM -GIUD GRAND -CU

FIGURE 26



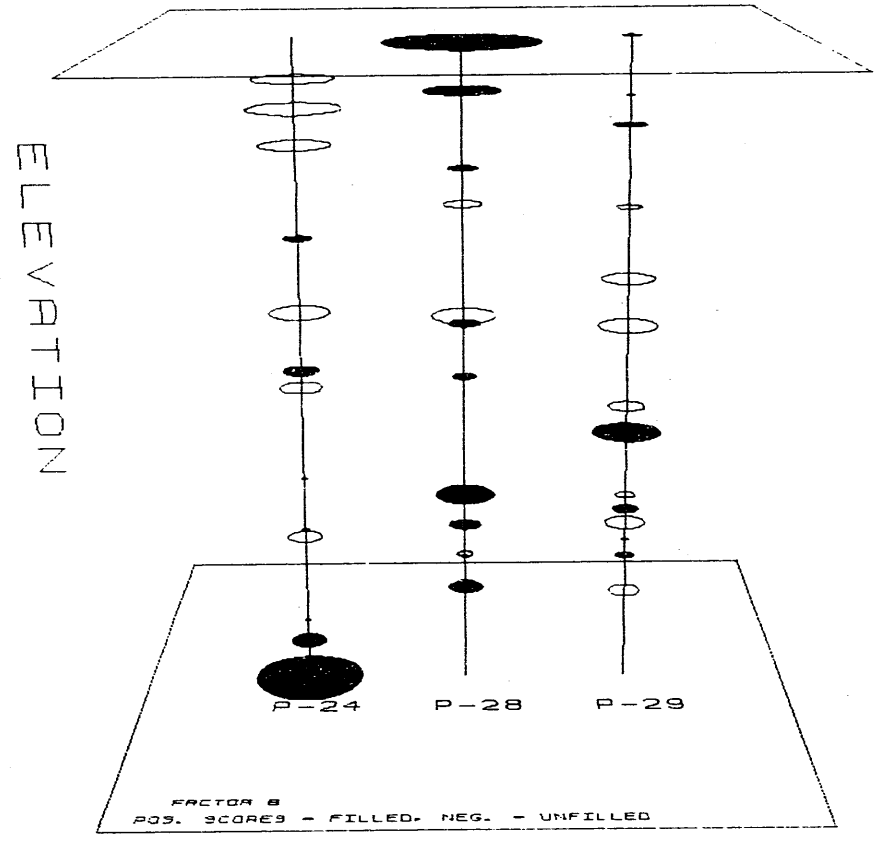
FUNGA 3ACC PH P3YC DIC

FIGURE 27



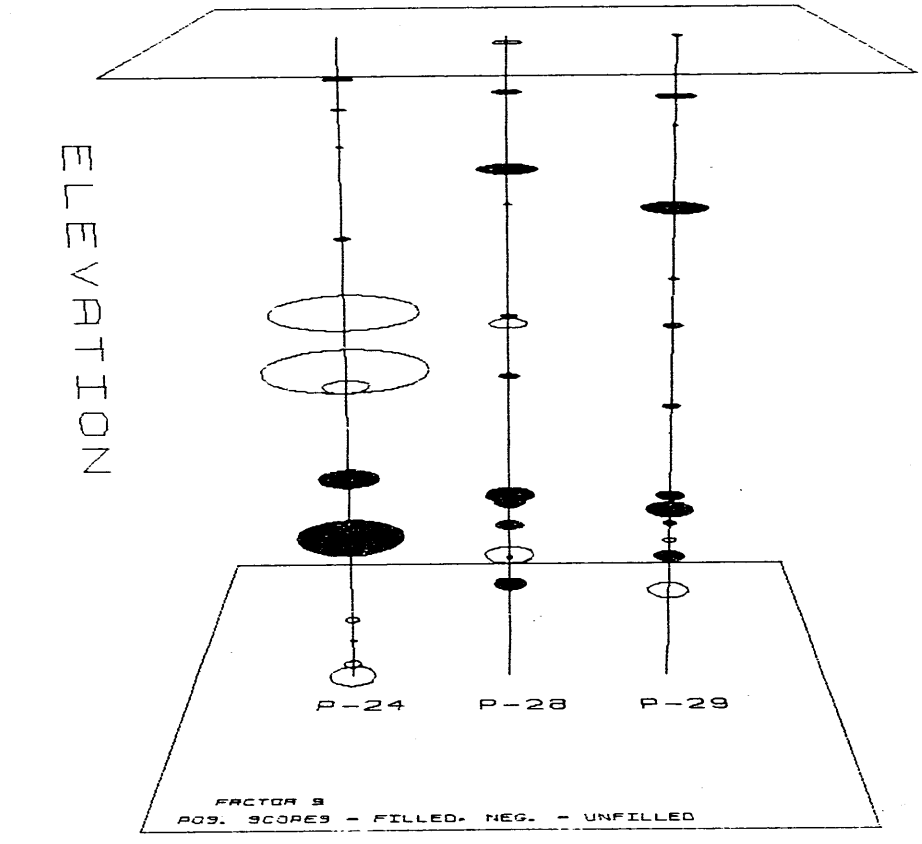
ADOC FACT 8

FIGURE 28



BA NOZI 9PA9 NA DOC CL

FIGURE 29



-CAT -C9:48 -CH4

APPENDIX A

SCATTER PLOTS

While simple measurement on measurement comparisons are useful, it is often valuable to find relationships among the measurements, or **correlations.** When we are examining the relationship between two variables it is often very useful to make a picture of the data so that we can see the relationship directly. One way of doing this is to use a pair of coordinate axes to represent the two variables and show each object (sample) as a point located by its value on the two variables. Such a graph is called a **scatter plot.** By examining the pattern of the plotted points we may make a judgement about the **relationship** between the variables. If the points scatter in a completely random fashion we say that there is no apparent relationship between the variables. The two variables have a completely random association and it is not possible to predict one variable from the other. However, when the scatter in the points appears elliptical we say that there is a relationship between the variables. If there is a consistent tendency for large values of one variable to be associated with large values of the other variable we say that there is a **positive** relationship. If large values of one variable tend to be associated with small values of the other variable we say that there is a **negative** relationship. When the points tend to fall into such an elliptical pattern we say that there is a **linear** relationship, reflecting the fact that the points tend to scatter about a straight line running through the ellipse. In cases where the points do not form a regular ellipse and do not fall around a straight line, we say that the relationship is **nonlinear.**

CORRELATION COEFFICIENTS

While examining the qualitative aspects of scatter plots is highly informative, it is useful to obtain a quantitative measure of the degree of the relationship between two variables, the **correlation coefficient,** r. (Most introductory statistics texts provide a formula for computing the correlation coefficient from the raw data. I shall not deal with it here.)

One way to interpret the magnitude of r is in terms of **"exceptions"** to the prevailing trend or pattern of the correlation. When r is close to +1.0 or -1.0, there are only minor exceptions to the trend of the correlation. When r is closer to zero, there are serious exceptions to the trend.

Another way of interpreting the magnitude of the r value is related to our earlier description of the scatter plot patterns. The r value magnitude is a measure of the "ellipticity" of the scatter. Values very close to unity indicate that the ellipse is very thin, infinitely thin in the case of a perfect correlation. A correlation coefficient of zero indicates that the ellipse has equal major and minor axes, the distribution pattern of the scatter plot is totally random, circular.

Some literature assigns adjectives to the numerical descriptors of correlation magnitudes. While such assignments are somewhat subjective, they are listed below for familiarization.

r value	Adjective Description
+/- .9	High correlation
+/- .7	Substantial correlation
+/- .5	Moderate correlation
+/- .3	Low correlation

Note that the magnitude of the correlation is determined by the scatter in the data. The significance is dependent on the magnitude of r and on the number of observations. Thus, it

is possible that a high correlation may not be statistically significant. (Consider the case of two points. The "correlation coefficient" is 1.0, but it has zero degrees of freedom and is, therefore, statistically not significant.) Correlations coefficients computed on about 40 observations must be greater than about 0.4 in order to be significant at the 99% confidence level.

APPENDIX 3

Investigator Update Comments Regarding Future Chemical Measurements

SAVANNAH RIVER PROJECT
EXPLORATORY DEEP PROBE

Exploratory Data Analysis

Investigator Update

center for environmental sciences

Laboratory for Chemometrics

Univeristy of Colorado at Denver

COMMENTS REGARDING FUTURE CHEMICAL MEASUREMENTS:
From: R. Meglen

I have examined the existing pore water chemistry data from P-24, P-28, P-29 with the objective of reducing the analytical chemistry load for the planned fourth hole. Keeping in mind that the primary objective of the fourth hole is to provide validation of the findings obtained from the first three holes, I believe that it would not be wise to drastically reduce the chemistry measurement suite. However, based upon exploratory statistical analysis of the existing data some modification may be justified.

I have concluded that two measurement characteristics make a variable a candidate for deletion; one, the chemical species is present at or below current analytical detection limits, two, the same chemical species is being measured by two different methods.

I have examined the chemical data base for variables which contain little information (i.e. few samples with a quantifiable amount of the chemical). The variables which fall into this category are indicated with a "DELETE" in the attached table. Some variables such as Mo, Ni, Co, Al also have few samples that are above detection limit; however, the measurement precision for these chemical species is sufficiently good that samples with quantifiable amounts are revealing useful information about the subtle differences among pore waters. This suggests that they will continue to be valuable in characterizing stratigraphic and spatial differences among cores. Therefore, for the fourth hole I would recommend retaining them in the analytical protocol.

Several chemical species were wisely determined by two independent methods. These species are prone to various analytical interferences that can introduce measurement bias. Having established the response/performance characteristics of these analytical methods it would be prudent to re-evaluate the need for continued measurement redundancy. I have indicated candidates for deletion because of redundancy by "SELECT" in the attached table.

I wish to emphasize that my criteria for deletion from future measurement protocols is based on statistical estimates only. It is clear that experimental design considerations ought to be the final basis for deleting any variables. There may be good chemical and microbiological justification for continuing to measure some of these. In addition, I think that Tom Garland's expertise regarding the analytical measurebility and cost/benefit evaluation should be the determining factor in making a final selection of a reduced measurement suite.

One particular cost savings appears to be available in deleting the ferrous/ferric iron measurement. While individual measurement of these species initially may have been justified on the basis of geochemical interests in the REDOX status of the pore waters, it is interesting to note that Eh measurements are most strongly correlated with nitrate (colorim), nitrate (ion chrom) and nitrite (colorim). Since the Eh measurement is not strongly correlated with iron (II) and iron (III) and both of these variables are strongly correlated with total iron, one could delete iron speciation from consideration and retain only the measurement of total iron. This conclusion is consistent with the factor analysis which shows that all three iron measurements load on the same factor (Factor #1). In this case one may interpret this as an indication that there is iron measurement redundancy.

Correlation Coefficients

	Fe	Fe2+	Fe3+	Eh
Fe	1.0			
Fe2+	.969	1.0		
Fe3+	.928	.936	1.0	
Eh	-.108	-.094	-.227	1.0

Similar observations regarding the nitrate and nitrite measurements may be made. There appears to be a significant chemical relationship between nitrate, nitrite and Eh. Colorimetric and ion chromatograph nitrate appear to be correlated highly indicating similarity of measurements. However, only colorimetric nitrite (NO2) shows a moderate correspondence with nitrate and Eh. In addition, the factor analysis indicates a separate factor for NO2 (ion chrom) Br and DOC (Factor #8). This factor probably represents an analytical interference among these species, not a significant geochemical/equilibrium relationship. Therefore, I would be inclined to drop ion chromatographic analysis for nitrite. I see no strong indication about which NO3 measurement would be preferable. I would leave that to Garland's judgement.

Correlation Coefficients

	NO3C	NO3I	NO2C	NO2I	Eh
NO3C	1.0				
NO3I	.884	1.0			
NO2C	.574	.533	1.0		
NO2I	-.008	-.042	.548	1.0	
Eh	.432	.480	.068	-.140	1.0

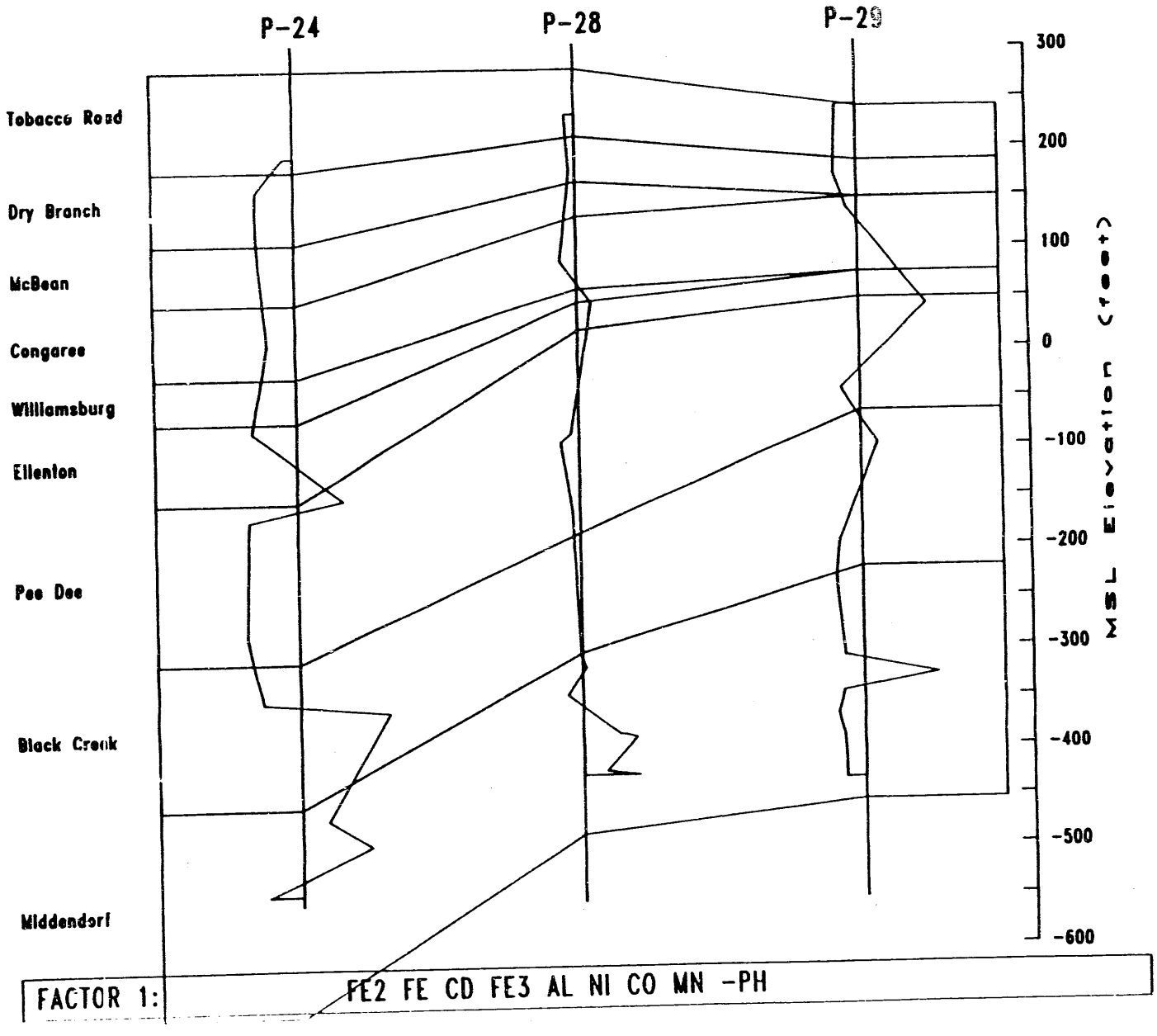
RECOMMENDATIONS FOR MODIFICATION OF CHEMICAL VARIABLES LIST

(R. Meglen)

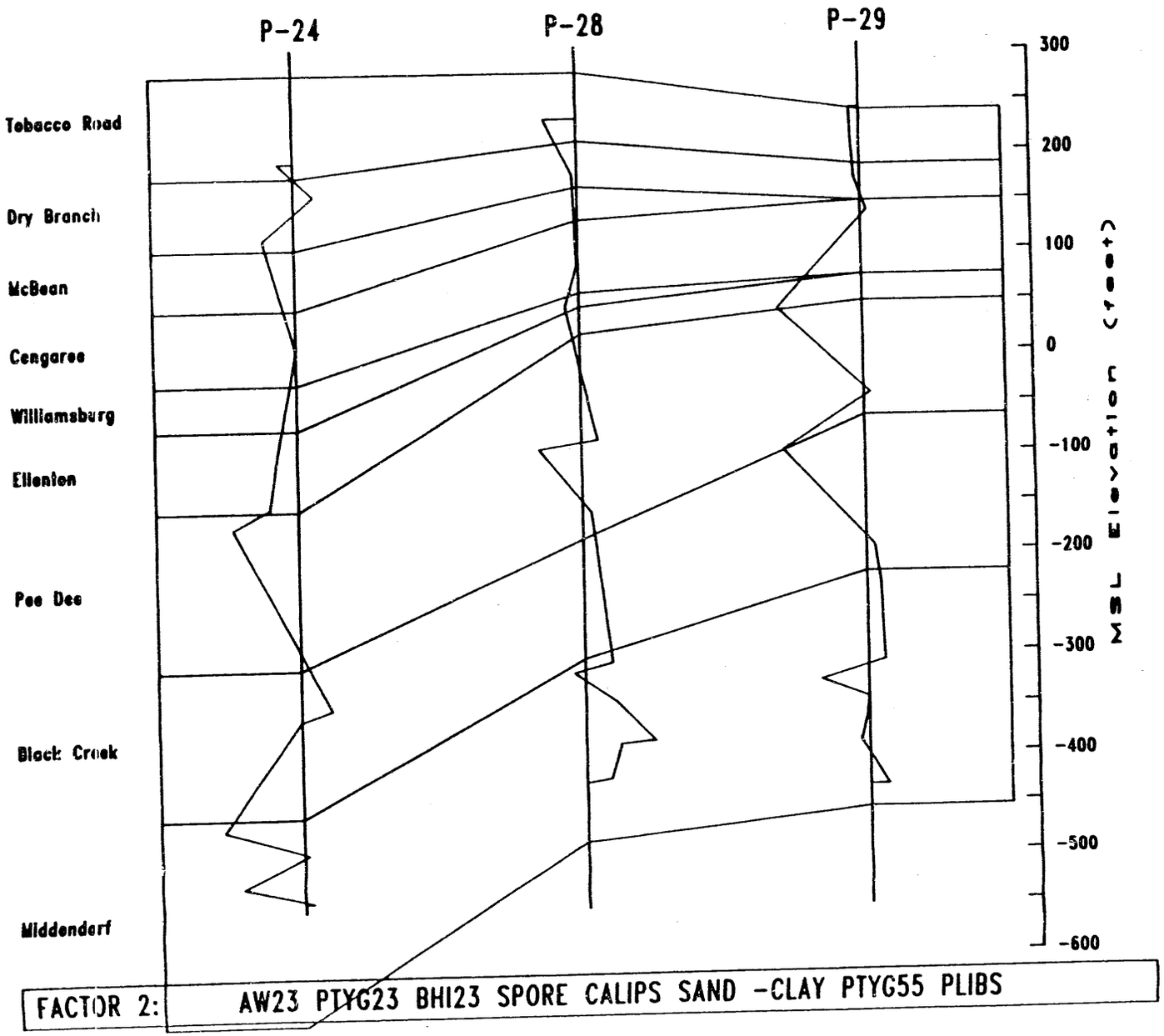
VARIABLE	RECOMMENDATION	COMMENT
VOL	KEEP	Volume in Fraction
INIM	KEEP	Initial Moisture
INCM	KEEP	Incubation Moisture
COND	KEEP	Conductivity (LOG)
PH	KEEP	Log Hydrogen ion activity
EH	KEEP	REDOX Potential
DC	KEEP	Dissolved Organic Carbon (LOG)
DIC	KEEP	Dissolved Inorganic Carbon (LOG)
NH4	KEEP	Ammonium (LOG)
NO3C	Select	Nitrate Colorimetric (LOG)
NO3I	Select	Nitrate Ion Chhromatography (LOG)
NO2C	Select	Nitrite Colorimetric (LOG)
NO2I	Select	Nitrite Ion Chromatography (LOG)
F#E	Select	Fluoride Selective Electrode (LOG)
F#I	Select	Fluoride Ion Chromatography (LOG)
CL	KEEP	(LOG)
BR	KEEP	(LOG)
PO4	KEEP	(LOG)
SO4	KEEP	(LOG)
FE3	Select	Ferric iron (LOG)
FE2	Select	Ferrous iron (LOG)
FE	Select	Total iron (LOG)
AL	KEEP	(LOG)
B	KEEP	(LOG)
BA	KEEP	(LOG)
CA	KEEP	(LOG)
CD	Delete	(LOG)
CO	KEEP	(LOG)
CR	Delete	(LOG)
CU	KEEP	(LOG)
K	KEEP	(LOG)
LI	KEEP	(LOG)
MG	KEEP	(LOG)
MN	KEEP	(LOG)
MO	KEEP	(LOG)
NA	KEEP	(LOG)
NI	KEEP	(LOG)
P	Delete	(LOG)
PB	KEEP	(LOG)
SB	Delete	(LOG)
SI	KEEP	(LOG)
SR	KEEP	(LOG)
TE	Delete	(LOG)
TI	Delete	(LOG)
ZN	KEEP	(LOG)

	FACTORS	NAME
1	Fe2 Fe Cd Fe3 Al Ni Co Mn -pH	Trans. Elements
2	AW23 PTYG23 BHI23 Spore CALIPS Sand -Clay PTYG55 PLIPS	Microbiol. A
3	Cond Sr SO4 DIC F#I F#E Mg Ca Ba	Major Elements
4	TMPN HALO TPC Morph TYEG -Mo Olig FungG PPC	Microbiol. B
5	-NO3C -NO3I -Eh NO2C Te	Nitrate/REDOX
6	FungP Coli -Pb Phos Psyc	Microbiol. C
7	-INCM -INIM -Cu -Clay Sand	Moisture
8	Br NO2I DOC -Vol	Analytical
9	Ti CH4 C3H8 Cat	Methane/Propane

SAVANNAH RIVER EXPLORATORY DEEP PROBE

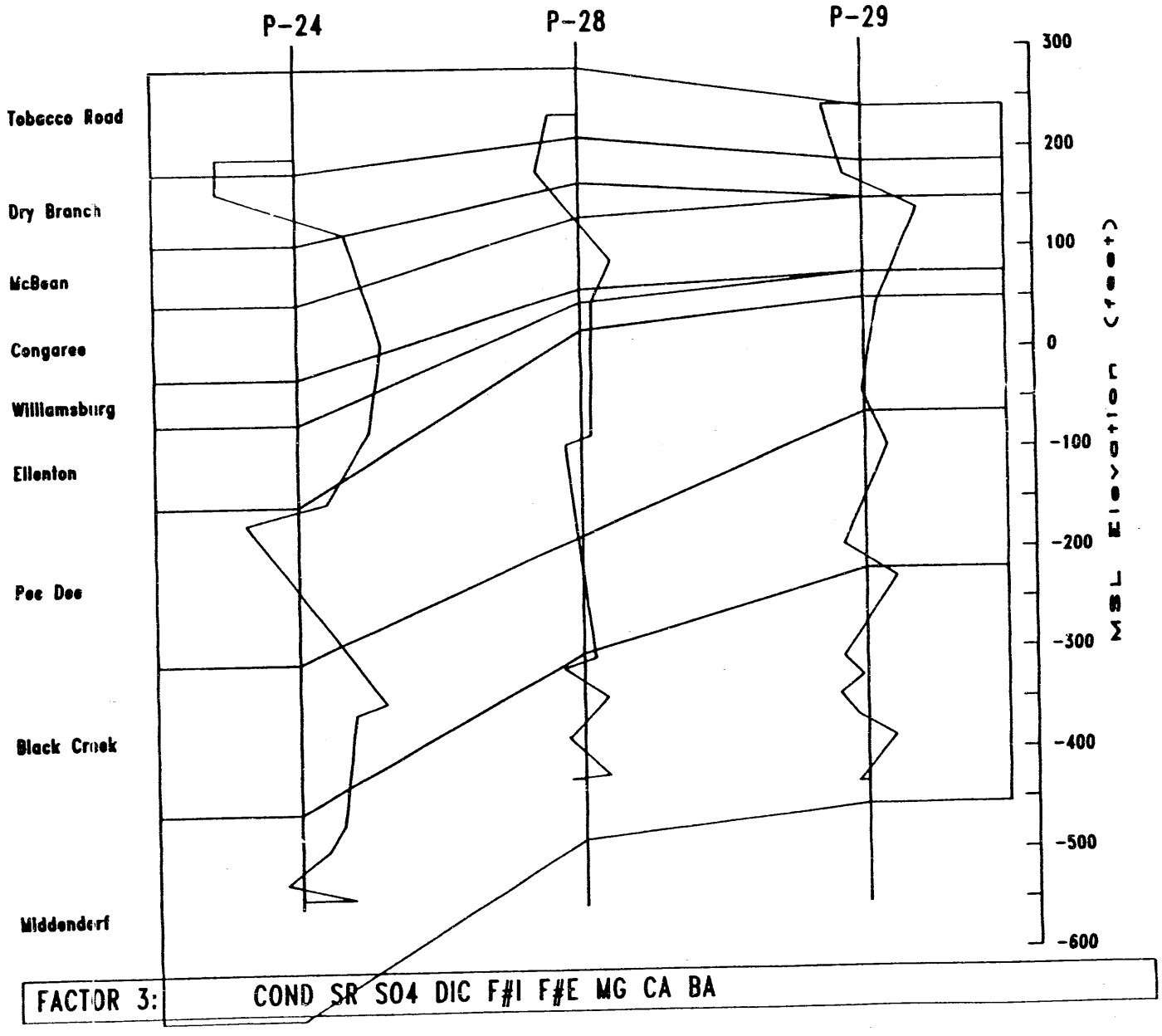


SAVANNAH RIVER EXPLORATORY DEEP PROBE

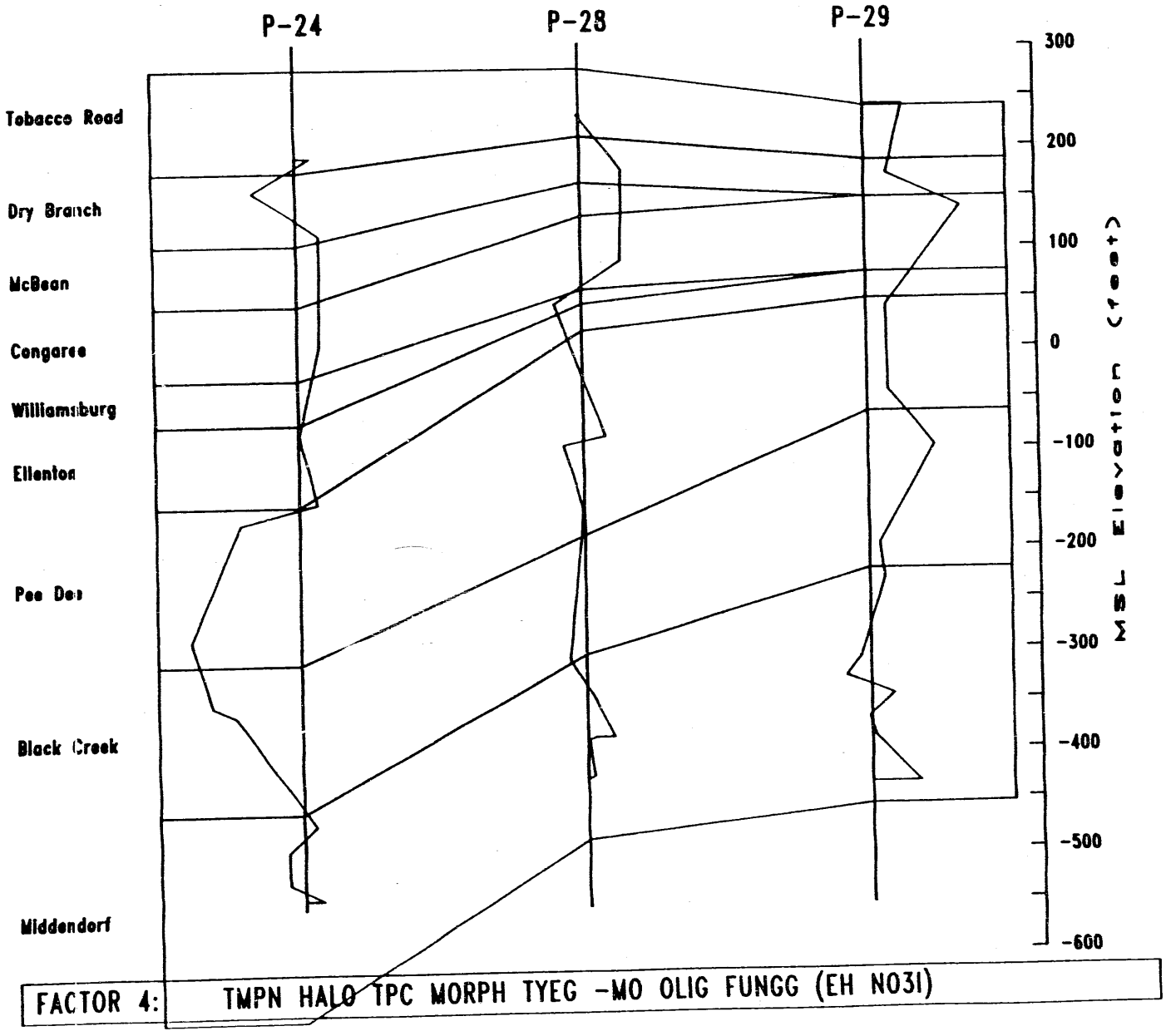


FACTOR 2: AW23 PTYG23 BHI23 SPORE CALIPS SAND -CLAY PTYG55 PLIBS

SAVANNAH RIVER EXPLORATORY DEEP PROBE

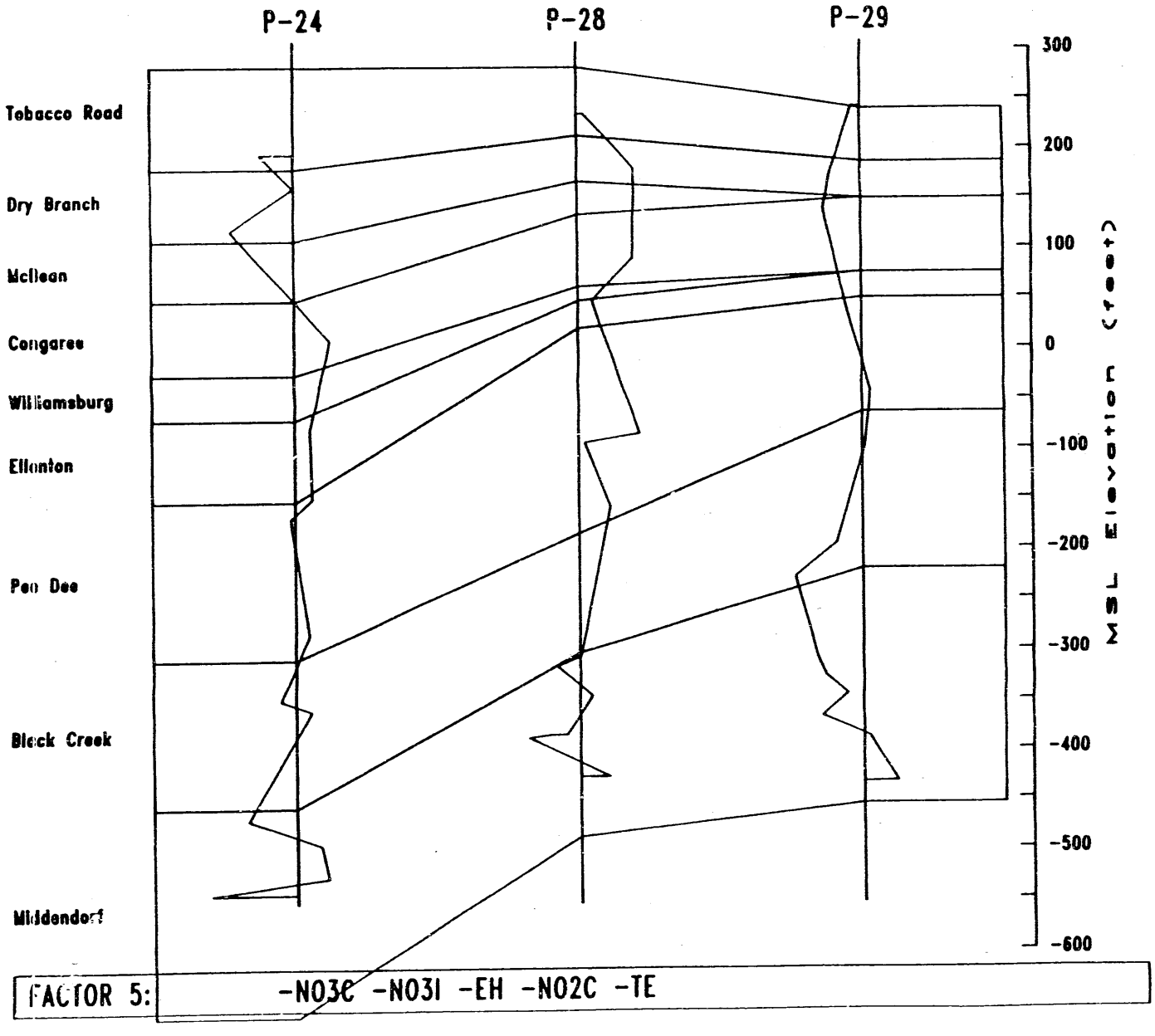


SAVANNAH RIVER EXPLORATORY DEEP PROBE



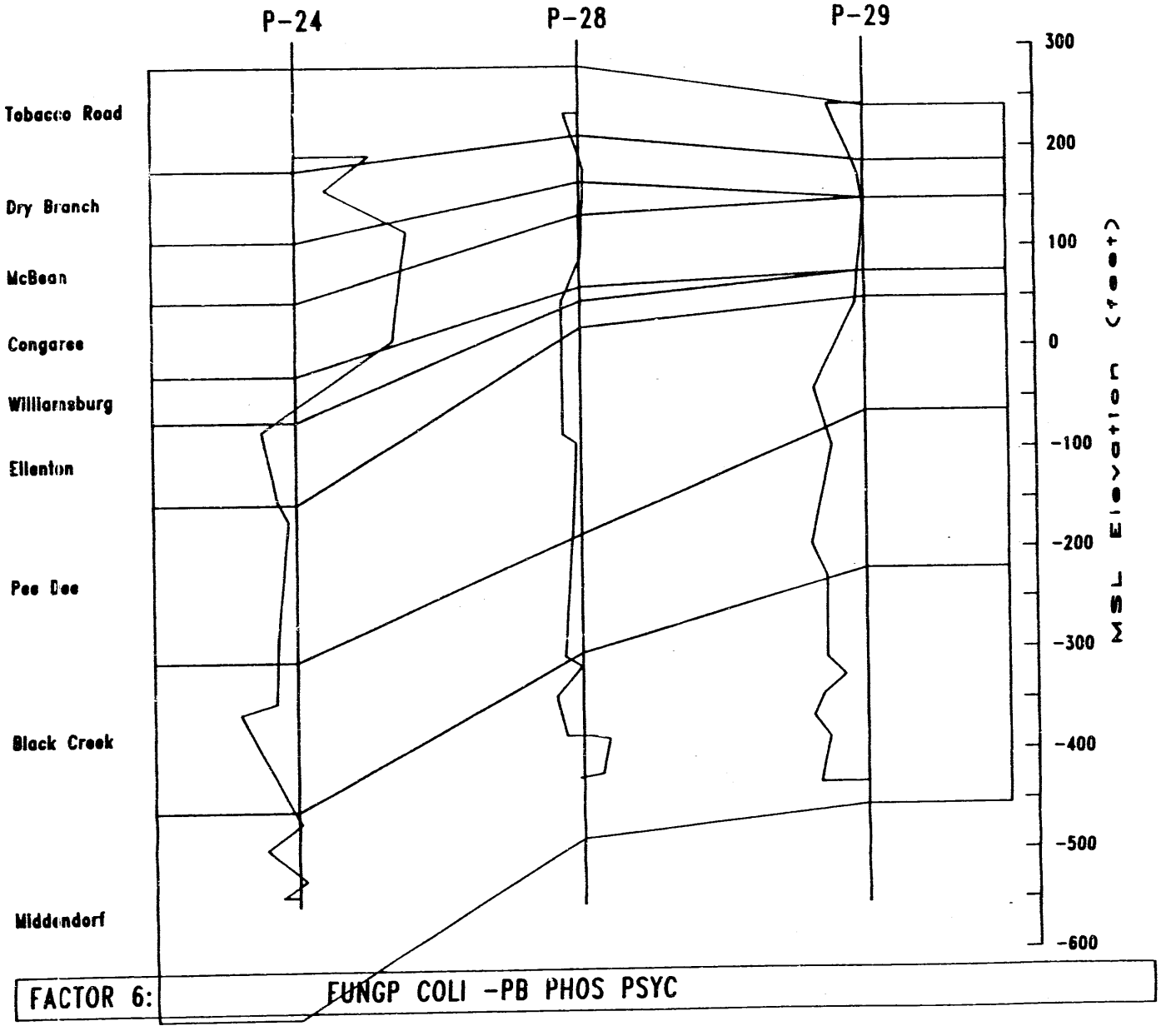
FACTOR 4: TMPN HALO TPC MORPH TYEG -MO OLIG FUNGG (EH NO31)

SAVANNAH RIVER EXPLORATORY DEEP PROBE



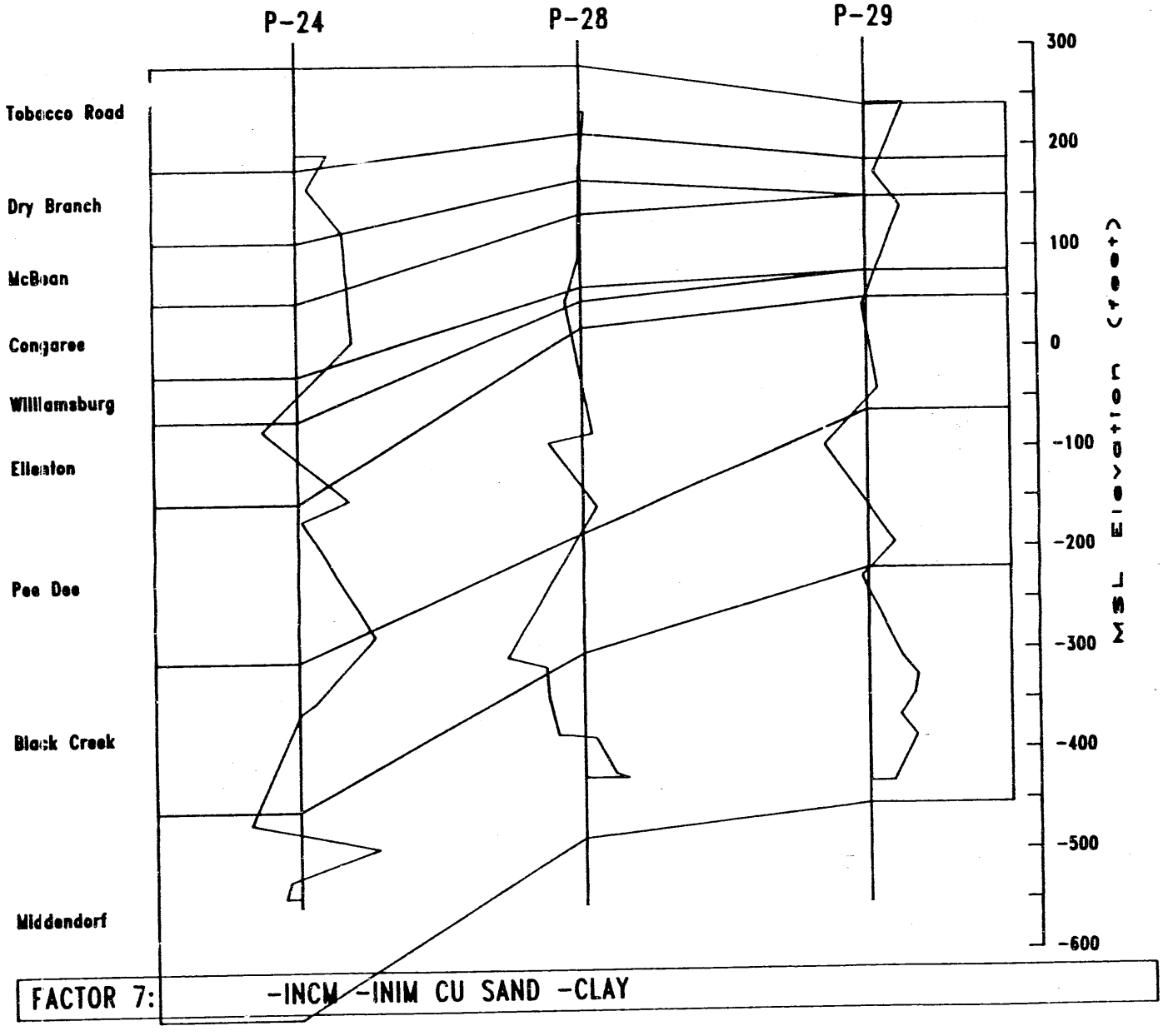
FACTOR 5: -NO3C -NO3I -EH -NO2C -TE

SAVANNAH RIVER EXPLORATORY DEEP PROBE

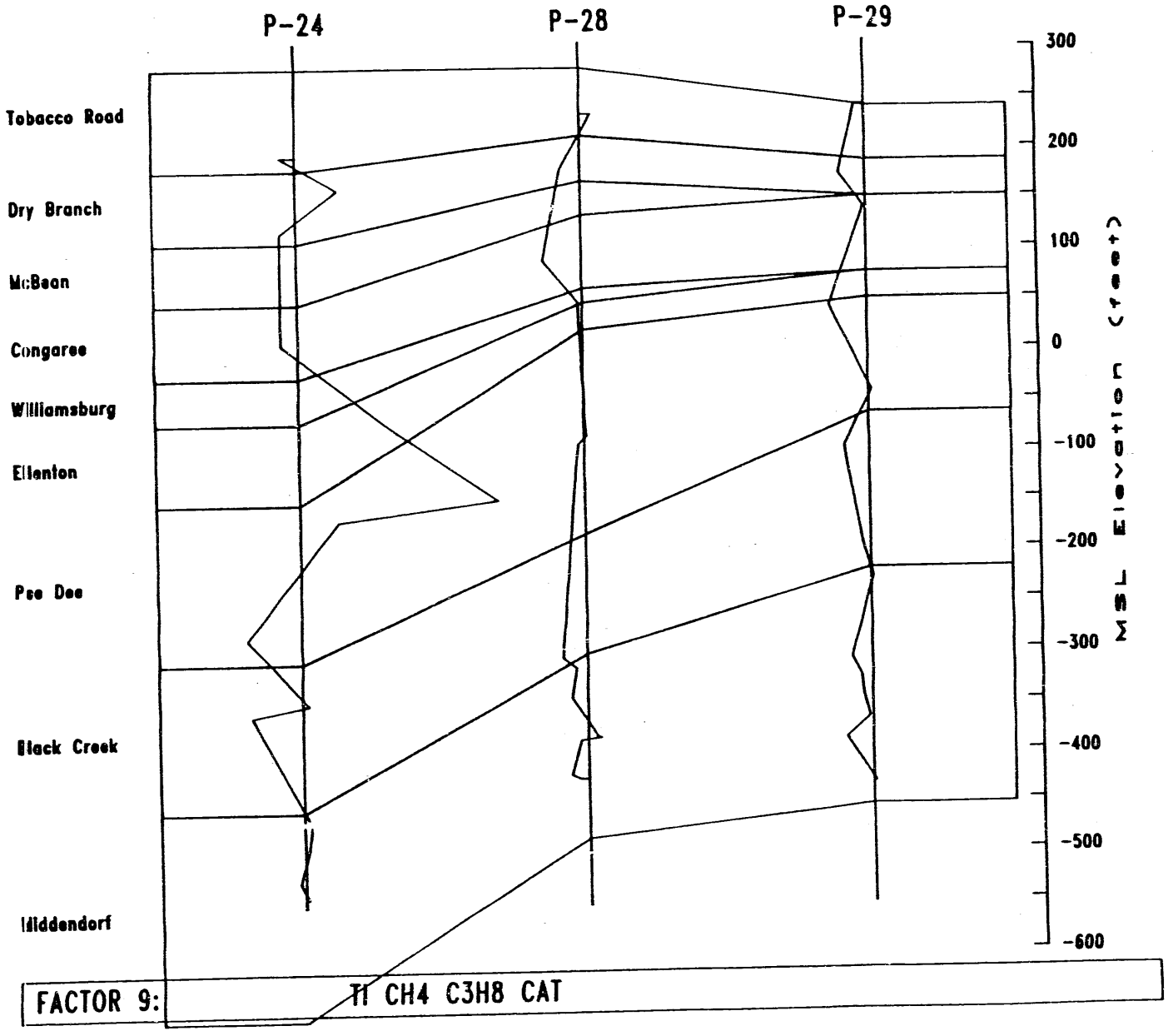


FACTOR 6: FUNGP COLI -PB PHOS PSYC

SAVANNAH RIVER EXPLORATORY DEEP PROBE



SAVANNAH RIVER EXPLORATORY DEEP PROBE



APPENDIX 4

Information Extraction and Collaborator Interaction in Interdisciplinary Studies

Information Extraction and Collaborator Interaction in Interdisciplinary Studies

Robert R. Meglen

Center for Environmental Sciences, Laboratory for Chemometrics
University of Colorado at Denver, 1200 Larimer St.,Box 136,Denver, CO 80204

ABSTRACT

The diversity of scientific disciplines represented in studies of the subsurface environment reflects the complexity of the system under study. The strength of these studies comes from the collaborative effort of hydrologists, geologist, chemists, and microbiologists. However, the experimental designs for these studies tend to focus on the sampling and measurement process. The majority of resources are focused on instrumentation and acquiring data, and too little effort is given to designing the experiment for optimal information extraction. When large multi-variable databases are generated and when interdisciplinary collaboration is required, a formal data analysis plan becomes a crucial element of experimental design. The answers will not just flow out the measurements. The database is a domain that requires probes and tools to extract relevant information. Instruments of reasoning to probe the data are as important to information extraction as measurement instruments are to data acquisition. Even the most experienced investigator may not have access to some of the sophisticated mathematical and statistical tools that may be necessary to fully exploit the data. Secondly, project success depends on a sustained collaborative interaction among the investigators. Past experience indicates that investigator collaboration is at a maximum in the sampling and initial data gathering phases of most interdisciplinary efforts. However, in later phases individuals tend to narrow their focus to their specific experimental objectives. Collaboration often is induced only by the external stimulus of report writing deadlines. A formal mechanism for effecting investigator interaction and information exchange should be incorporated into the data analysis plan. Most projects could be strengthened by the participation of a data analysis specialist. The data specialist provides technical expertise needed to critically evaluate data, functions as a conduit for its transfer, and acts as a catalyst for collaborative interaction.

INTRODUCTION

The results from innovative subsurface microbiology research programs indicate that new data on the nature of the link between the geosphere and biosphere are being generated. The diversity of the scientific disciplines represented in these subsurface microbiology projects reflects the complexity of the system under study. The research being carried out by national laboratory and university research scientists is addressing fundamental questions about the abundance of microorgansims and factors controlling microbial activity in the complex subsurface hydrologic and geochemical environment. Long term implications of this research for mitigating contamination are clear and researchers share the broader objective of linking the basic science with applied work. Data on a large number of variables are already being acquired. However, massive quantities of data alone will not ensure that an adequate mechanistic description will obtain. Data are not information. Achieving this objective requires a formal data analysis plan to exploit the existing data base and to permit integration of the many-variable data into a useful information resource.

Past experience obtained from large multidisciplinary studies indicates that project success is often limited because experimental designs for these studies tend to focus on the sampling and measurement process. The majority of resources are focused on instrumentatic and acquiring data, and too little little effort is given to designing the experiment for optimal information extraction. When large multi-variable databases are generated and when interdisciplinary collaboration is required, a formal data analysis plan becomes a crucial element of experimental design. The answers will not just flow out the measurements. The database is a domain that requires probes and tools to extract relevant information. Instruments of reasoning to probe the data are as important to information extraction as measurement instruments are to data acquisition. Even the most experienced investigator may not have access to some of the sophisticated mathematical and statistical tools that may be necessary to fully exploit the data. Secondly, project success depends on a sustained collaborative interaction among the investigators. Past experience indicates that investigator collaboration is at a maximum in the sampling and initial data gathering phases of most interdisciplinary efforts. However, in later phases individuals tend to narrow their focus to their specific experimental objectives. Collaboration often is induced only by the external stimulus of report writing deadlines. A formal mechanism for effecting investigator interaction and information exchange should be incorporated into the data analysis plan. Most projects could be strengthened by the participation of a data analysis specialist. The data specialist provides technical expertise needed to critically evaluate data, functions as a conduit for its transfer, and acts as a catalyst for collaborative interaction.

Most large-scale research programs are not limited in their ability to generate valid data. In fact, they are often characterized as "data-rich". However, many projects have data handling and interpretation weaknesses that restrict data sharing among the participants. Five key generalizations about large-scale multidisciplinary projects are outlined here.

Too little attention is given to experimental design.

We tend to focus on the measurement process and undertake massive efforts to design and assemble the necessary instrumentation. Too little effort is given to designing the experiment for optimal information extraction. When we have a good understanding of the phenomena under study we tend to select the correct variables. However, when confronted with systems having complex variable interactions about which we have inadequate theory to guide us, we tend to measure inefficiently and often our designs are insufficiently flexible to allow for the discovery of unexpected relationships. Without consciously designing for the unexpected the experimental success is jeopardized. Before acquiring any new data a formal statement of the specific measurement objectives needs to be made.

Historical data are underutilized.

While most experiments use past experimental results to formulate test hypotheses and to help in the design of new explorations, historical data are seldom formally incorporated in the interpretive phase of data analysis. Clearly, data obtained with less sophisticated instrumentation than currently available, or data obtained with different intent, may not be applicable in the current context. However, we are often too quick to dismiss historical data from active participation in the analysis. Powerful multivariate statistical techniques are available to uncover systematic bias and to assess validity. These techniques can dramatically improve our ability to cut through the fuzzy facts and focus on the essential features that strengthen our inferences.

Too little attention is given to data analysis plans.

Formal data analysis plans are seldom included as an integral part of research plans. It is assumed that the investigator has the expertise to analyze the data generated from his/her own experiment. However, when many variables are present and when interdisciplinary projects are undertaken, it is necessary in recognize that the answers may not just flow out the measurements. A formal plan for converting the assembled data into information is as important to the problem-solving endeavor as experimental design is to the measurement process. Unanticipated system complexity may require interpretive tools to which even the most expert investigator may not have access. The data base may be viewed as a domain that requires probes and tools to extract relevant information. Just like the measurement process itself, appropriate instruments of reasoning need to be assembled if data are to be fully exploited. We must be sensitive to the investigator's desire to examine and interpret their own data. Indeed, they are uniquely capable of interpreting the results of their own experimental design. However, as the problems become more complex, it becomes increasingly important to call upon the expertise of data analysis specialists. It is important that this collaboration be established prior to data acquisition. Calling in an expert afterwards may be too late.

Mechanisms for interdisciplinary interaction are seldom formalized.

Past experience indicates that investigator collaboration is at a maximum in the sampling and initial data gathering phases of most interdisciplinary efforts. However, in later phases individuals tend to narrow their focus to their specific experimental objectives. Collaboration often is induced only by the external stimulus of report writing deadlines. A sustained immersion into the interdisciplinary context facilitates interpretive insights. The mechanism for effecting this interaction is improved if a formal information exchange is incorporated into the data analysis plan. The information/data analysis specialist can be an effective conduit for this transfer and can act as a catalyst for sustained interaction.

Data and information management is viewed too narrowly.

Too often when people talk about data base management systems they think in terms of hardware; computers, terminals, graphics displays, etc. A well designed information management system should be more than just computer software or a commercial data management system. It should incorporate human creative elements using interpretive aids to display the data conveniently, summarize its information, induce thinking about its content, and facilitate its use as an instrument of reasoning. The data specialist is a key individual in the interdisciplinary team, not just a librarian. That individual should have the technical expertise needed to critically evaluate the data and function as an aid to analysis as well as to retrieval. All projects benefit from the broader perspective of a competent generalist. Thus, one central role for information management is to provide day-to-day continuity that implements the data analysis plan and keeps it productive.

END

5-20-91