CORRES CONTROL
OUTGOING LTR NO
DOE ORDER #

94 RF 02971

DIST	LTR	ENC
AMARAL M E		
BEEMAN H S		
BRANCH D B		
CARNIVAL G J		
COPP R D		
DAVIS J G		
FERRERA D W		
HANNI B J	X	
HARMAN L K		
HEALY T J		
HEDAHL T		
HILBIG J G		
HUTCHINGS N M	X	
KELL R E		
KIRBY W A		
KUESTER A W		
MAHAFFEY J W		
MANN H P	X	
MARX G E		
McDONALD M M		
McKENNA F G		
MONTROSE J K		
MORGAN R V		
POTTER G L		
RIZZUTO V M		
RISING T L		
SANDLIN N B		
SETLOCK G H		
STEWART D L		
STIGER S G		
SULLIVAN M T		
SWANSON E R		
WILKINSON R B	X	
WILSON J M		
WYANT R D		
J K Nowens	X	
S P Nixeneg	X	X
R S Roberts	X	X
M Anderson	X	
W S Busby	X	
A L Pomrose	X	X
T P O'Rourke	X	X
CORRES CONTROL	X	X
DOCU RECORD/080	X	
RAFFC		
WATS/T130G		

CLASSIFICATION

UCNI		
UNCLASSIFIED		
CONFIDENTIAL		
SECRET		

AUTHORIZED CLASSIFICATION
UNCLASSIFIED
CLASSIFICATION WAIVER PER
CLASSIFICATION OFFICER
DATE

IN REPLY TO RFP CC NO

ACTION ITEM STATUS
PARTIAL/OPEN
CLOSED
FOR APPROVALS

ORIG & TYPIST INITIALS
SJO / mp

EG&G ROCKY FLATS

EG&G ROCKY FLATS, INC
ROCKY FLATS PLANT, P O BOX 464, GOLDEN, COLORADO 80402 0464 • (303) 966-7000

0010023613

94-RF-02971

FILE

March 14, 1994

Shirley J Olinger
Assistant Manager
Environmental Safety & Health
DOE, RFO

STATISTICAL METHODOLOGY FOR BACKGROUND COMPARISONS - SGS-179-94

Refs (a) S G Stiger ltr, SGS-126-94, to S J Olinger, Statistical Methodology, February 28, 1994
(b) S J Olinger ltr, (01234) to S G Stiger, Statistical Methodology, February 14, 1994

This letter responds to the above referenced (b) memorandum, using the deliverable date extension requested in reference (a)

Attached please find responses to comments made in reference (b) This letter has four attachments Attachment (a) responds to comments made in reference (b), and attachment (b) contains responses to previous Environmental Protection Agency and Colorado Department of Health comments, which addressed Dr Gilbert's methodology Attachment (c) responds to agency comments on the site-to-background methodology, and attachment (d) contains the statistical site-to-background methodology itself Attachments (b), (c), and (d) have been modified per DOE comments in Reference (b)

If you have questions or comments, please contact Richard S Roberts at extension 8508

S G Stiger
Associate General Manager
Environmental Restoration Management

RSR mp

Attachments
AsStated (4)

Orig and 1 cc - S J Olinger

cc
E A Howard - DOE, RFO
M H McBride - " "
J M Roberson - " "
R. J. Schassburger - " "
M N Silverman - " "
L. W Smith - " "
B K Thatcher - " "

ADMIN RECORD

SW-A-003592

1/32

4/32

ATTACHMENT A

RESPONSE TO DOE-RFO COMMENTS ATTACHED TO DOE MEMO ER:BKT:01234
DATED FEBRUARY 14, 1994

2

DOE/RFO COMMENTS ON RESPONSES ATTACHED TO SGS-667-93 ATTACHMENT A

1. EPA Specific Comment 1 - You have not responded to their suggestion that the same field sampling and laboratory procedures be used for both background and site data. DOE/RFO agrees with this comment. This is truly a background comparison issue. Provide an accurate and appropriate response to this comment.

Concur. The same procedures were used for field sampling and laboratory procedures for both site and background. The response to EPA Specific Comment 1 has been changed to state this.

2. EPA Specific Comment 2 - Essential nutrients have not been eliminated from the protocol in the statistical methodology. State this fact in your response. Also state that EPA withdrew this comment at our September 29, 1993 meeting.

Concur. The response to EPA Specific Comment 2 has been rewritten to incorporate this.

3. EPA Specific Comment 6 - EPA, CDH, and DOE/RFO agreed at our September 29, 1993 meeting that DQOs were an important issue, but should be dealt with independently from the statistical methodology. State this in your response.

Concur. The response to EPA Specific Comment 6 has been modified to incorporate DOE's comment.

4. EPA Implementation Issue 3 - See DOE/RFO comment immediately above regarding DQOs. Restate here.

Concur. The response to EPA Implementation Issue 3 has been modified to state this.

5. EPA Implementation Issue 5c - This conflicts with the response to EPA Implementation Issue 1. Eliminate the inconsistency in both the responses to comments and in the statistical methodology document. There is confusion regarding detection vs. reporting limits.

Concur. The replaced value equals 0.5 times the reported detection limit. The methodology and the response to EPA Implementation Issue 5c have been changed to state this. There are several different detection limits: instrument detection limit, contract required detection limit, etc. The reported detection limit is available from the RFEDS data base.

6. You have not responded to EPA's general comments in their September 21, 1993 letter to DOE/RFO. Provide written responses to their general comments.

3

Concur. These have been added to the beginning of Attachment B.

7. CDH Comment 3 - See DOE/RFO comment 5 above. Be consistent.

Concur. The replaced value equals 0.5 times the reported detection limit. The methodology and the response to EPA Implementation Issue 5c and CDH comment 3 have been changed to state this.

DOE/RFO COMMENTS ON RESPONSES ATTACHED TO SGS-667-93 ATTACHMENT B

1. CDH Comment 9 - Your response is counter to prior written direction from DOE/RFO. Move the Preliminary Exploratory Data Appraisal to the Data Presentation Section as requested by CDH. A meeting is not appropriate as this issue has been previously discussed between EG&G and DOE/RFO.

Concur. The document has been changed.

2. CDH Comment 10 - Your response is counter to prior written direction from DOE/RFO. State in your response and in the statistical methodology document that this information will be informally discussed with EPA and CDH at a meeting with DOE/RFO. We do not have to commit to a formal written deliverable. A meeting is not appropriate as this issue has been previously discussed between EG&G and DOE/RFO.

Concur. The document has been changed to incorporate this.

DOE/RFO COMMENTS ON ATTACHMENT C

1. Figure 1-2 - This figure refers to how the 1993 Background Report proposed that these comparisons be made and is not discussed in the methodology document. Thus, we request that this figure be deleted from the methodology document.

Concur. Figure 1-2 has been deleted. In addition, the paragraph that referenced it (page 6, immediately following the section on the t-test) has been deleted.

2. The first paragraph of page 2 states that the background data sets will be taken from the 1993 Background Geochemical Report. However, the surficial soil data from Rock Creek and the associated UTLs were not included in this report. In addition, no provision is made for supplementing these data with the planned background surficial soil sampling for FY 94. The text should be corrected to reflect these items.

Concur. The text on page 2 has been rewritten to reflect this.

3. The first paragraph under "Data Collection and Validation" on page 2 states that data will be used for OU comparisons without waiting for 100% validation. It further states that the impacts of using non-validated data will be discussed on a case-by-case basis. This may result in a complete rerun of the statistical comparison of background and RFI/RI data if only a few percent of the data are rejected in the validation process. The individual OU Workplans, QAPjP, and QA Workplan addenda should be reviewed regarding the use of rejected data. The methodology should state that the OU Workplan, QAPjP, and QA addenda will be reviewed prior to using rejected data.

Clarification. The section on page 2 has been rewritten for greater clarity. Non-validated data will be used only for draft RFI/RIs. Final RFI/RI reports will use only data that have undergone validation. Data that have been rejected will not be used.

4. The last sentence on page 5 states that a discussion of detection limits will be given, but this discussion was not included. We request that this discussion be provided.

Clarification. The intent was not that the discussion be included in the Guide, but that it be provided in the RFI/RI report that would follow this guidance. The statement on page 5 has been rewritten to clarify this.

5. All figures in the statistical methodology document should have both figure numbers and consistent captions. Correct this situation.

Concur. This has been done.

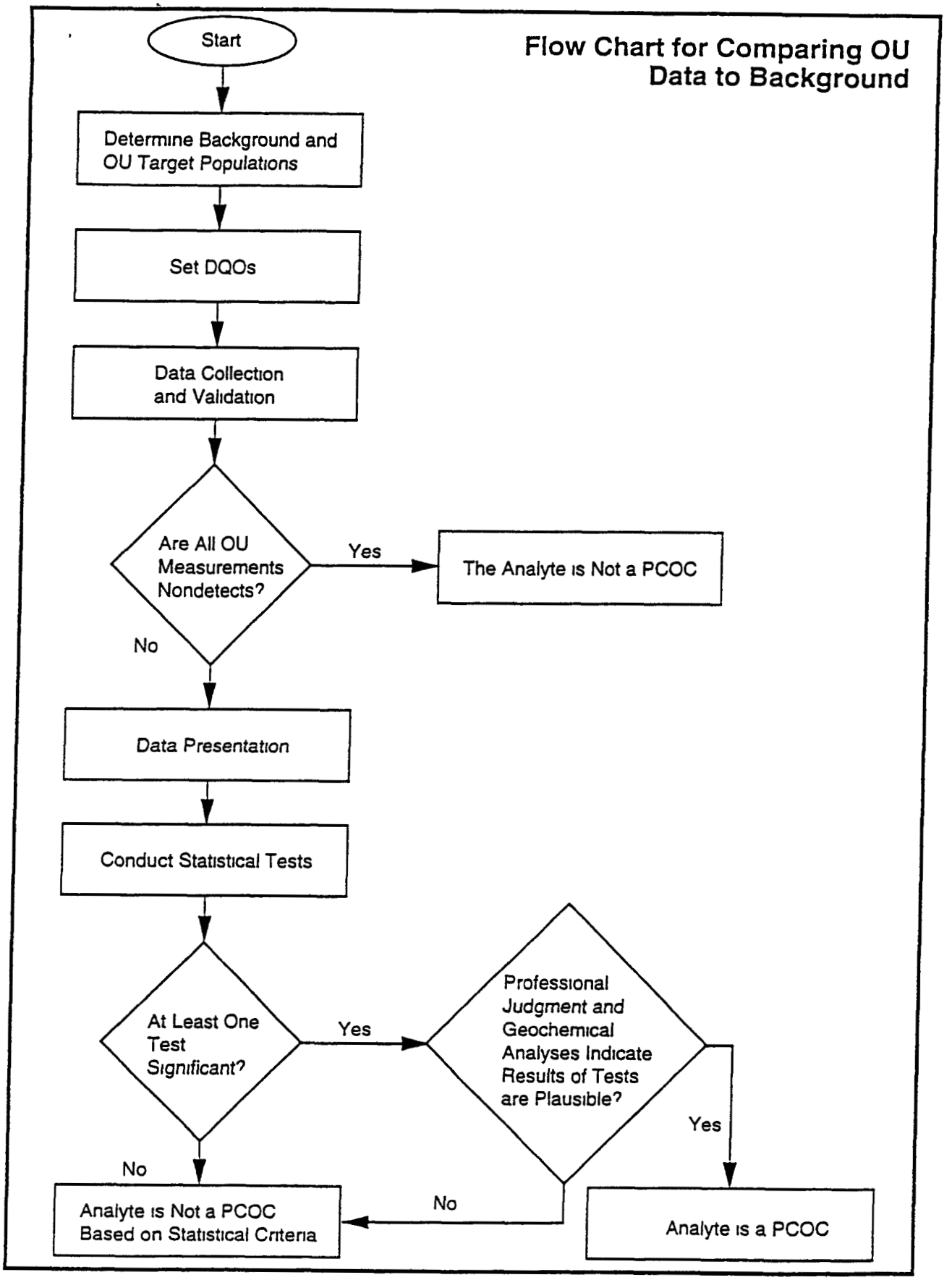

Figure 1
Flow Chart for Comparing
OU Data to Background

GENERAL APPROACH TO DETERMINING "CONTAMINANTS"

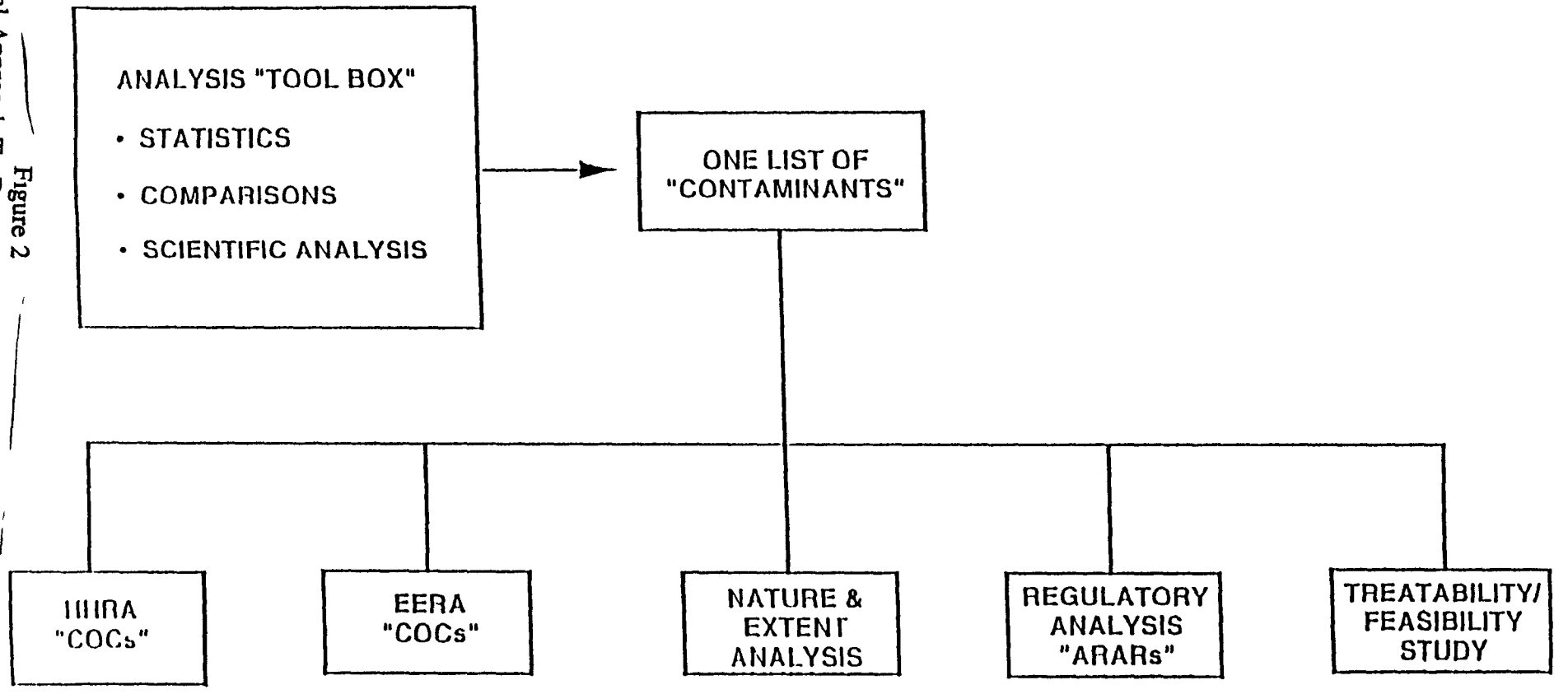

Figure 2
General Approach To Determining "Contaminants"

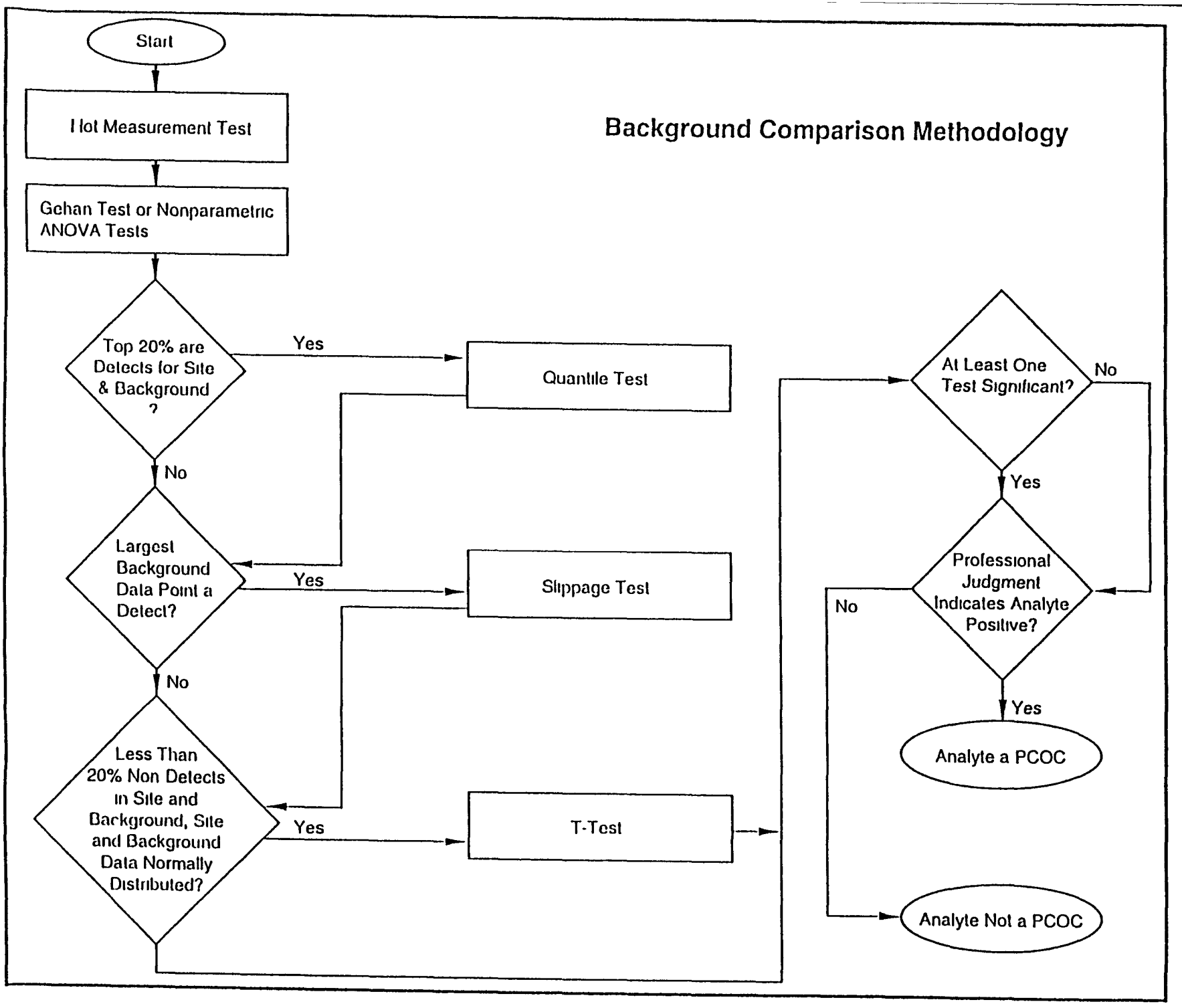

Figure 3
Background Comparison Methodology

ATTACHMENT B

RESPONSES TO EPA LETTER 8HWM-FF - STATISTICAL COMPARISONS TO BACKGROUND AT ROCKY FLATS DATED SEPTEMBER 21, 1993 AND TO CDH LETTER - STATISTICAL METHODS FOR THE COMPARISON OF REMEDIAL INVESTIGATION DATA TO BACKGROUND DATA AT ROCKY FLATS PLANT, DATED SEPTEMBER 13, 1991

RESPONSES TO EPA LETTER 8HWM-FF - STATISTICAL COMPARISONS TO BACKGROUND AT ROCKY FLATS DATED SEPTEMBER 21, 1993:

GENERAL COMMENTS

1. Overall, the report is outstanding. It succinctly outlines a comprehensive paradigm for the background analysis of inorganic chemicals at RFP. It is obvious that the multitiered approach, incorporating specific data quality objectives, presentation and graphic analysis, and a series of six statistical tests has been well thought-out and all possible scenarios considered and problems anticipated. It directly addresses the predominant contentious and divisive issue, the proper application of the upper tolerance limit (UTL) approach that has been advanced by DOE.

On a purely technical level, the approach is well-balanced. However, the report appears to be overly concerned with Type I or false positive errors and not as concerned with Type II or false negative errors. From a risk assessment standpoint, a Type I error can be easily managed if it is unknowingly included in the risk assessment since the analysis can be revisited and professional judgement applied if the risk associated with the chemical in question becomes unacceptable. In contrast, a Type II error cannot be so easily managed. If a Type II error is made, the chemical will be incorrectly eliminated early in the COC selection process and will not be further considered Although it is desirable to minimize or eliminate both types of errors from the analysis, from a public health perspective it is preferable to make a Type I error. Chemicals included in the risk assessment from a Type I error will not automatically be remediated. EPA recommends that for risk assessment, sampling design should specify the probability of a Type I error as 20% and the probability of a Type II error as 10% or less. This is an important item to reach consensus on between EPA, CDH, and DOE.

> **Clarification.** It is necessary to reach a compromise between acceptable Type I error rates, acceptable Type II error rates, and cost. Each OU should have addressed these issues in the OU-specific workplan approved by DOE and the Agencies. The corresponding background data are now a matter of historical record, as this sampling program was terminated after four years of data collection. At the September 29, 1993 meeting, all parties agreed to the Type I error rates incorporated into the current plan, which is 5% for the Gehan, quantile, slippage, t-test, and Wilcoxon Rank Sum tests, with the Type II error rate left unspecified. However, it should be noted that the actual Type II error rate has been reduced, and the Type I error rate has been increased, because of the battery of tests. The UTL test will increase the Type I error rate and reduce the Type II error rate as well.

2. One additional problem that is not addressed in Dr. Gilbert's report, perhaps because it was outside the scope of work, involves data aggregation. This is a fundamental issue that has yet to receive the proper amount of focused attention. Without an established methodology for aggregating data within different environmental media, the time and effort expended in executing the sophisticated statistical approach presented in this report will be misspent. Although the report touches on some aspects of this broad problem, it does not directly discuss the issue.

Therefore, EPA, CDH, and DOE need to address it.

Clarification. Data aggregation is another topic, being addressed by DOE/RFO, CDH, and EPA separately from this forum, which deals strictly with site-to-background comparison.

3. If the agencies can agree that the above concerns will be addressed, the background analysis approached developed by Dr. Gilbert provides a well-balanced methodology that will, if implemented properly, lead to a robust background analysis. This objective, scientific approach will result in verifiable conclusions, expedite the review and comment period, and prevent an overreliance on professional judgement.

No response necessary.

SPECIFIC COMMENTS

1 Page 2 Seventh Bullet It is suggested that the same field sampling and laboratory procedures be used for both background and site data. The statement should be extended to include data aggregation. Past review of RFP data from operable units showed inconsistencies in the methodology used to aggregate data. Problems encountered at this phase will be magnified at later stages of the background analysis.

Concur. The same field sampling and laboratory procedures were used for both background and site data.

Clarification. Data aggregation is another topic, being addressed by DOE/RFO, CDH, and EPA separately from this forum, which deals strictly with site-to-background comparison.

2. Page 4, Task 1, Observation 1, Third Bullet This statement suggests that background analysis should be the initial state in selecting COCs. This is consistent with the COC selection methodology developed for Rocky Flats by DOE, EPA, and CDH. However, in order to manage DOE's effort in background comparisons, we point out that it is not necessary to carry all chemicals through an elaborate, time consuming statistical analysis if they can be eliminated as essential nutrients or as infrequently detected chemicals. It may be more cost-effective and expeditious to simply eliminate chemicals on the basis of these two preliminary criteria than to conduct a background analysis only to eliminate them later based on the background analysis. We suggest that DOE consider this in the development of a plan to implement Dr. Gilbert's approach.

Clarification. Essential nutrients have not been eliminated from the protocol in the statistical methodology. This comment was withdrawn by EPA at the September 29, 1993 meeting.

3. Page 5, Task 1, Observation 4, Second Bullet This statement expresses concern about measurements that are less than the contract required detection limits (CRQL) but above instrument detection limits (IDL). According to Risk Assessment Guidance for Superfund,

Human Health Evaluation Manual, Volume I, Part A, these measurements should be "J" coded and interpreted as estimated values They should not be viewed as non-detected chemicals. If they are currently classified as non-detect chemicals in the RFP background geochemical report, the entire validation process currently in place should be reevaluated.

Clarification. There has been confusion over the detection limits and their application A qualifier of "J" indicates that the reported value is between the instrument detection limits and the contract required detection limits. A non-detect has a reported value of a detection limit, not the detected value, and conveys less information than a "J".

4. Page 9, Paragraphs 3 and 4 The essence of this discussion is that a hot measurement (HM) concentration should serve as a "safety net" that can prevent "hot spots" from passing unnoticed in a risk assessment. It should be noted that this need has been previously recognized and was addressed in the original flow chart devised during the summer 1992 meetings involving EPA, DOE, and CDH. At that time, it was agreed that a risk-based concentration (RBC) would effectively serve as the "hot measurement." Although a UTL has some utility in identifying hot spots, there is no need to conduct a lengthy analysis if the highest detected concentrations do not exceed a predetermined RBC and pose an unacceptable human health risks. Thus, it is possible to have measurements above the UTL but below an RBC in which case there would be little reason to consider the chemical further.

Clarification. The Guide for Conducting Statistical Comparisons of RFI/RI Data and Background Data at the Rocky Flats Plant (called The Guide subsequently) addresses statistical determination of the presence or absence of analytes, and does not address human health effects. For each OU, additional tests will determine if the analyte concentrations present are below regulatory (ARARs) and/or human health effect (PRGs) levels, but that is external to the statistical discussion at hand.

5. Page 10, Third and Fourth Bullet. This statement refers to lowering the potential for a Type I, false positive error to using a 99 percent UTL on the 99 percentile. However, this concern is not properly balanced against the potential for a Type II error. A false negative could have profound consequences on the risk assessment and subsequent remedy selected for the site.

Do not concur. If the 95% UTL were used, then a very high percentage of data points would be considered pCoCs, because theoretically, even a background population will have 5% of readings above the UTL. A site, even if its concentration levels are slightly above background, may have considerably more than 5% of its readings above the $UTL_{95/95}$. Any analytes that show a false negative on this test will still be considered pCoCs if they test positive on any of the other statistical tests.

6. Page 11, Second Paragraph This paragraph suggests that data quality objectives (DQOs) be established at the design stage of the studies. Although this is a relevant comment in the context of planning a background analysis, the background and most of the OU planning and sampling has already been completed. Thus, this comment is appropriate in theory but there is little chance for implementation. Revitalized effort should be directed to establishing DQOs where they were not previously established, and analyzing whether the sampling efforts

IMPLEMENTATION ISSUES

1 EPA, DOE, and CDH must reach consensus on procedures for defining non-detects.

Concur. The Guide states that non-detects will be considered to be one-half of the reported detection limit, in accordance with EPA guidance.

2 EPA, DOE, and CDH must reach consensus on what hot measurement value should be used.

Concur. Our methodology uses a value of $UTL_{99/99}$

3. EPA, DOE, and CDH must establish data quality objectives which address acceptable power and confidence levels, required detection limits, and anticipated data aggregation.

Clarification. The draft RIs for each OU have a section for reviewing data quality Each OU manager bears the responsibility for ensuring that DQOs are met for his or her OU. This issue will be dealt with independently from the statistical methodology, as was agreed to by EPA, CDH, and DOE at the September 29, 1993 meeting.

4 EPA, DOE, and CDH must revisit the assumptions which Dr Gilbert lists on page two of his cover letter Are these assumptions valid? What are the consequences if the assumptions are violated? Can this be handled in an uncertainty analysis?

Clarification. All of the assumptions listed, except for the last four, are difficult to quantify and are thus not "valid" or "invalid" These last four are now answered individually.

The same field-sampling techniques are used for background and site, so this assumption is valid.

Measurements are not always validated by third-party subcontractors before the draft RFI/RI statistical testing has been completed, so this assumption is not valid. When the data validation results have been obtained, the data are reanalyzed, and the final RFI/RI contains no rejected data.

Background data were checked for outliers, per EPA comments upon the 1992 Background Geochemical Report, and extreme outliers were excluded from statistical analysis in the 1993 Background Geochemical Report, so this assumption is not entirely valid. However, OU data outliers are not typically deleted, although data from the OUs are checked for "geochemical reasonableness", and any unusual results are discussed in the ensuing reports.

The instrument detection limits are not always reported in the data bases, so this assumption is not completely valid. However, the costs of recovering this information would be considerable.

5. EPA, DOE, and CDH must reach consensus on a paradigm for implementation. The issues to be worked out include:

a. The appropriate background data sets by analyte, medium, and location.

Concur. The section of The Guide entitled "Determine Background and OU Target Populations" addresses how this will be done.

b How to deal with clearly non-random (e g , spatial) patterns.

Concur. The Guide states in the Professional Judgement section that spatial patterns are subject to professional judgement, which is then subject to EPA and CDH review

c. Measurement errors and multiple non-detects.

Concur. Measurement errors are an inevitable part of physical data. Efforts are taken throughout the data-collection process to minimize errors. When non-detect replacement is necessary (i.e., for t-tests or UTL tests), non-detects are dealt with by replacing the data value with ½ of the reported detection limit.

d Structure for the formal statistical tests

Concur. The Guide furnishes this structure

e. Data aggregation for comparison in the statistical tests

Clarification. Data aggregation is another topic, being addressed by CDH and EPA separately from this forum, which deals strictly with site-to-background comparison.

RESPONSES TO CDH LETTER - STATISTICAL METHODS FOR THE COMPARISON OF REMEDIAL INVESTIGATION DATA TO BACKGROUND DATA AT ROCKY FLATS PLANT, DATED SEPTEMBER 13, 1991

1 The Division would like to emphasize the importance of effective graphical presentation of data to enhance the understanding and interpretation of the statistical tests. The Division believes that the development of effective graphical procedures to display and interpret both site and background data is essential to the usefulness of the methodology and should not be overlooked or down-played. The Division requests that specific graphical techniques be developed and included in the "statistical strawman" methodology.

Concur. The Guide specifically addresses graphical techniques.

2. The Division does not recommend the use of a risk based hot measurement comparison value in the hot measurement comparison. The use of risk based decisions is not appropriate in the context of comparisons to background

Concur. The hot-measurement comparison value is not risk-based.

3. As noted in Dr Gilbert's report, the proper treatment of non-detects and multiple detection limits is critical to the implementation of his recommendations. Both of these issues occur frequently in Rocky Flats data sets. Therefore, the Division recommends that DOE emphasize specific protocol for proper treatment of non-detects and multiple detection limits in the "strawman" methodology.

Concur. The Guide states that non-detects will be dealt with by replacing the data value with ½ of the reported detection limit.

4. The Division agrees with Dr. Gilbert that professional judgement is necessary in evaluating the results of statistical tests. However, it is not the Division's intention that professional judgement be a substitute for an inadequate site investigation or as a tool to dismiss dubious data. The scope of appropriate professional judgement and limitations on its application should be outlined in the "strawman" methodology Guidelines and criteria for making decision based on professional judgement should also be identified.

Concur. The Guide restricts professional judgement to several specific areas.

ATTACHMENT C

RESPONSES TO EPA: HESTMARK LETTER 8HWM-FF RECEIVED 10/25/93 AND TO CDH LETTER "DOE PROPOSED METHODOLOGY FOR STATISTICAL COMPARISON OF REMEDIAL INVESTIGATION DATA AT THE ROCKY FLATS PLANT" FROM G BAUGHMAN TO R. SCHASSBURGER, DATED 10/13/93

Response to EPA Hestmark letter 8HWM-FF received 10/25/93

1 To determine the appropriate background and operable unit populations for comparison, we understand that some matching of the two populations is done by geologists and chemists. Data for an analyte in a non-background area are grouped according to a combination of background classes which represent independent background populations. A table that cross references the operable unit populations and the background populations will be provided.

Concur. The strawman has been changed to require tables that cross-reference OU media to background media.

2 A more explicit statement of the null hypothesis that is being tested will be included In addition, a fixed p value of 0 05 will be used for each of the inferential statistical tests as written in the strawman proposal There was some inconsistency in what was written in the proposal and what was stated in the meeting regarding the p value A fixed value of 0 05 is what we will accept

Concur The strawman states that p values must be less than or equal to 0 05 to demonstrate a significant difference from background Footnote 3 on page 5 of the strawman, which was not clear on this point, has been deleted

3 All references to comparison of background and operable unit populations for organics will be removed Background comparisons apply to inorganics and radionuclides only

Do not concur. Although background comparisons for organics are not commonly used, there are instances when it may be applicable, in which wide-ranging organic contamination is due to non-site-specific anthropogenic sources. We want to retain the option of performing background comparisons for these organics, when geochemists or geologists determine that it is applicable to do so In these instances, we will retain the burden of proof, and the applicability of the comparison will be subject to EPA and CDH approval.

The strawman has been rewritten to state that background comparisons for organics will be done on a limited, case-by-case basis, subject to EPA and CDH approval

4 The use of professional judgement in interpreting the results of the graphical displays and statistical analyses will be limited to consideration of spatial distribution, temporal distribution, and pattern recognition concepts. The strawman proposal included five additional criteria These will be deleted in the final implementation document.

Concur The five criteria (intermedia interactions and geochemical processes, not an expected contaminant, blank data, regional background range, and influence of field activities) have been deleted.

5 The non-background population is defined as the entire operable unit remedial investigation set. The data aggregation for the purpose of background comparison will be done within the area defined by the operable unit boundaries.

Concur Analysis will be done on an OU-wide basis

6 The attached flowchart, "Background Comparison Methodology", distributed at the meeting will be clarified. It is EPA's understanding that all the data sets will undergo the hot measurement test and the battery of inferential statistical tests (Gehan, Quantile, Slippage, and T-Test) provided the data satisfies the conditions stated in the strawman and on the flowchart. If any one of these tests, including the hot measurement test, shows significance, the analyte will be further considered, using professional judgement, as a contaminant of concern. The flowchart would benefit from the addition of decision blocks after each test indicating the next step if significance is demonstrated or not.

Clarification The chart "Background Comparison Methodology" attached to EPA's memo is not the same as that distributed at the September 29, 1993 meeting and contained within the strawman proposal The difference is that nonparametric ANOVA tests are given as options to the Gehan test in the chart within the strawman proposal Because the Gehan method is not standard and will therefore incur practical liabilities (e g , the method has not been adequately tested and verified, preliminary usage shows it to require excessive man-hours, and subcontractors will need to be instructed in its use), we want to retain the option of performing standard nonparametric ANOVA testing, using the Wilcoxon or Kruskal-Wallis tests, instead of the Gehan test.

Additional clarification. The suggested decision blocks are not necessary. All tests will be performed, if applicable, regardless of whether other tests demonstrate significance.

Concur with the need to redo the flowchart This has been done.

6. (continued) We also have some specific questions that need to be addressed in the final document:

a. What happens to data which is carried through the slippage test but does not qualify for the t-test?

Clarification. The data that do not qualify for the t-test will be routed to the "At Least One Test Significant?" block The flowchart has been revised to show this.

b What is the basis for the 20% detect value as the criteria for the Quantile test? How does this criteria relate to the criteria for applying this test as stated in Dr. Gilbert's report on page 20?

Clarification. Dr. Gilbert's method proposed looking up tabulated values for n and r parameters. The quantile test could be correctly applied only if the largest n values were all detects. Our statisticians have stated that, typically, this restriction equates to the

largest 20% or less of the combined sample sizes being detects, and recommend using a flat 20% to simplify application.

c What is the basis for the criteria of N>20 value for background and operable unit data?

Clarification. Our statisticians derived this value from application of the Central Limit Theorem for a two sample problem. If both samples have N=20, then there will be 38 total degrees of freedom, which will permit assumptions about the distribution.

7 EG&G's claim that these impacts [of implementing Dr Gilbert's recommendations] could range from $30,000 up to $120,000 per operable unit is not supported by the information provided. In fact, it appears that there is some evidence that implementation will not negatively impact costs or schedules.

Do not concur. Because the Gilbert method requires additional work, there will be cost and/or schedule impacts.

In addition to the impacts mentioned above, cost impacts may result if the Gehan method is used. For OU11, approximately 200 hours were required to perform the Gehan test, when less than 40 hours would have been sufficient to perform standard ANOVA testing However, the majority of these costs appear to be one-time costs such as coding development Subsequent testing on the same OU indicate that the cost impacts may be as little as 30 hours for a small data set.

19

Response to CDH letter "DOE Proposed Methodology for Statistical Comparison of Remedial Investigation Data at the Rocky Flats Plant" from G. Baughman to R Schassburger, dated 10/13/93

1 To minimize any potential future misunderstandings of this agreement, the Division feels that it is critical for the Agencies to develop a formal guidance/policy document institutionalizing the agreement. The Strawman document was written for the purpose of facilitating agreement among the Agencies However, the end users of this document will be the operable unit managers and sub-contractors preparing and reviewing RFI/RI reports The majority of these people were not involved in the development of this methodology. It is critical to the future of this agreement that final documentation of this agreement be developed to clearly and concisely guide future end users in the implementation of this methodology This formal guidance should be completed in parallel with the implementation of the agreement

Concur When the strawman has been completed and accepted by all concerned parties, it will then be rewritten as a procedure for statistical comparison of OU data to background

2 The Division recommends that the title of this document be revised to more accurately reflect its content and intent, that being methodology and guidelines for the comparison of site data to background data. The Division proposes the title, "Guide for Conducting Statistical Comparisons of RFI/RI Data and Background Data at the Rocky Flats Plant," for consideration.

Concur The CDH's proposed title is an improvement to the current title, and has been adopted.

3. One of the central themes of Dr. Gilbert's recommendations was the need for statisticians to be involved throughout the entire process. However, statistician involvement is not discussed in the methodology. The division requests that the role of the statistician in implementation of this methodology be clarified in this document.

Concur. Statisticians will be employed to verify that the methods used are correct. The strawman has been rewritten to incorporate this.

4 The Division does not believe that references to specific DOE sub-contractors are appropriate in this document. The Division recommends DOE review all references to sub-contractors and, where appropriate, modify the reference to more accurately reflect DOE's role and responsibilities.

Concur. References to DOE subcontractors have been eliminated.

5 This section (Determine Background and OU Target Populations) outlines the steps for matching site and background populations. However, it is unclear exactly how the matching will be implemented. The Division recommends that the rationale for combining media/geology groupings for testing be detailed in this section. For example, any criteria for minimum group size necessary for statistical testing should be specified. The Division further recommends adding a table or diagram depicting the general rationale for grouping data by media and geology

Concur The strawman states that the OU will match one or more of several specified background media. In addition, the strawman has been changed to require that a cross-reference be performed between the site and one or more background media.

6 As discussed during the September 29th meeting, and emphasized by Dr. Gilbert, it is critical to statistical hypothesis testing that the hypothesis to be tested is explicitly defined and clearly stated. The Division recommends a statement of the test and null hypotheses, in both "english" (narrative qualitative description) and statistical terms, be added to this section of the methodology so there is no misunderstanding of what is being tested. This statement should also address confidence and power requirements for the tests

Concur The strawman has been modified to require statistical and prose statements of the null and alternative hypotheses

7. The Division does not agree with the blanket statement at the beginning of this discussion, "Under current IAG schedule conditions, analytical data will not be 'validated' when the background comparisons will be made in each draft report." This claim is not substantiated by the schedules submitted by DOE in the approved OU work plans and is in direct contradiction to Dr. Gilbert's Task 5 recommendations. Dr. Gilbert states that, "These data quality evaluations are conducted prior to descriptive graphical analyses and formal statistical tests." In finalizing this methodology, the Division recommends that DOE follow Dr. Gilbert's recommendations for data validation before formal graphical presentation and statistical testing The need for variance from this approach will be considered by the Division on an OU specific basis.

Do not concur Under the present system of data validation, the non-validated data are used only for the draft RFI/RI The final RFI/RI is based solely upon validated data. The lag time between receiving data from the laboratory, and validated data from the independent subcontractor can exceed one month. Waiting for 100% validation may impact schedules, but will probably not change the results in the final RFI/RI. The potential impacts of using non-validated data at each OU will be discussed on a case-by-case basis.

8. The Division recommends DOE add a discussion of detection limits to this section of the methodology In the past there has been confusion as to what detection limits are being reported and used (instrument detection limits vs contract limits vs reporting limits). Part of this confusion may be because detection limits have not been formal discussed. This section should state what detection limits are to be used in statistical testing and how they are

21

determined from the RFEDS data set.

Concur. The strawman addresses detection limits, and it specifies how determinations are made on how to handle non-detects

9 The Division recommends that this section (Preliminary Exploratory Data Appraisal) be moved to the Data Presentation section.

Concur. This section has been moved to the Data Presentation section.

10. The Division interprets this section as describing the informal data analysis conducted during RFI/RI preparation and not normally included in the formal RFI/RI report. The Division recommends adding language to indicate that this informal data analysis will be made available and reviewed with the regulators in evaluating the appropriateness of the scope of the formal RFI/RI proposal.

Clarification. We have added language to this section to clarify that this informal data analysis will be informally discussed with CDH, EPA, and DOE/RFO. However, this will not constitute a formal deliverable

11. The Division does not agree with DOE's recommendations that box plots are applicable only when there are no non-detects The problem of estimating percentiles for data sets with multiple non-detects was not resolved by Dr Gilbert. The Division recommends that when a reasonably small percentage of non-detects are present, percentiles be estimated using Maximum Likelihood Estimation (MLE) techniques in constructing box plots.

Concur. We will provide box plots unless the percentage of non-detects exceeds 50% The 50% figure is chosen for consistency with the 1993 Background Geochemical Characterization Report (September 30, 1993).

12. The Division does not agree with DOE's suggestion that histograms are not useful for small or highly censored data sets, such as inorganics. As stated by Dr Gilbert, such histograms are not likely to be useful in visually assessing whether the data sets are better modeled by a normal or lognormal distribution However, they may still be useful to visually compare the spread, central tendency, and skewness of the two data sets to look for differences that may be important.

Concur. We will provide histograms unless the percentage of non-detects exceeds 50%. Bars in the histogram will be shaded to indicate the percentage of detects and non-detects within each bar interval.

13. The Division recommends that a discussion be added to this section of the methodology to address what to do when a UTL 99/99 can not be reasonably estimated or is unknown (ie small or highly censored background data set).

Concur. We have modified the strawman to state that professional judgement, and use of

22

geochemical background data from the literature, will be used. The result will be a geochemical interpretation of data, subject to agency review and approval.

14 The reference in Footnote 2 to OU 1 is not appropriate and should be removed The inferential tests conducted at OU 1 were the result of a compromise agreement, are not precedent setting for other OUs and are not the tests being proposed in this document. However, as stated in this note, limited professional judgement as presented later in this document may be applicable.

Concur This footnote has been deleted

15. This discussion (Footnote 3) should be moved to the DQOs or statistical test definition section of the document.

Clarification. This footnote has been deleted We intend to use a p value of 0 05, and the footnote made that intent unclear

16 The Division does not agree with the limitations DOE has placed upon the Slippage Test. The slippage test can be applied to data sets when the largest background point is a non-detect. If the largest background data point is a non-detect then logic must be applied to determine if the slippage test is applicable, but the test should not be categorically eliminated

Concur We have rewritten the strawman to state that, if the largest background data point is a non-detect, we will apply judgement to investigate whether or not the slippage test is applicable

17. The Division recommends limiting the use of professional judgement to the first three criteria; spatial distribution, temporal distribution, and pattern recognition. In addition, it is recommended that the introduction to this section include acknowledgement that in applying professional judgement, the "burden of proof" lies solely on DOE. Professional judgement will only be considered by the Division on a limited basis where well documented and defensible evidence is presented

Concur We have eliminated the last five criteria from the strawman, and acknowledged that we will bear the burden of proof.

18 To make the process more efficient the task of eliminating non-detected analytes should be completed prior to data presentation. The flow chart should be modified to reflect this change.

Concur We have changed the flowchart. CDH's comment improved the process.

19 This flow chart is confusing and difficult to follow due to the many multiple and undefined branches. To minimize the potential for misunderstanding this chart must either be clarified or deleted.

23

Concur. The flowchart is too important to delete. It has been clarified Lines denoting the flow of information have been deleted, keeping only the lines denoting flow of control, in accordance with common flowcharting techniques Decision blocks have been transformed into diamond shapes Alternative "No" paths have been added for the blocks labeled "No Non-Detect Present ..OU Data Normally Distributed?", and "At Least One Test Significant?" Finally, the block representing the conditions which must be met prior to performing the t-test has been changed to reflect the conditions given in the text.

ATTACHMENT D

GUIDE FOR CONDUCTING STATISTICAL COMPARISONS OF RFI/RI DATA AND BACKGROUND DATA AT THE ROCKY FLATS PLANT

~~25~~

Guide for Conducting Statistical Comparisons of RFI/RI Data and Background Data At the Rocky Flats Plant

General

This document is intended to provide guidelines for OU-to-background comparisons of data, and to explicitly discuss approaches to the issue of determining OU-specific contamination. The OU-to-background comparison will be applied for inorganics and radionuclides. In addition, the comparison may occasionally be performed for organics on a limited, case-by-case basis, subject to EPA and CDH approval

It is important to establish a common approach leading to a common list of possible contaminants for each OU To this end, Figure 1, **GENERAL APPROACH TO DETERMINING "CONTAMINANTS"** was developed. In this general technique, a "Tool-Box" approach is employed to arrive at one common list of contaminants for each OU (or subdivision), for all functional aspects of the RFI/RI and CMS/FS

As indicated, several disciplines such as the Human Health or Ecological Risk Assessors and Regulatory specialists may pare the list of contaminants to "Contaminants of Concern" (COCs) based on factors germane to their application (e.g , toxicity)

The text below follows Figure 2, **FLOWCHART FOR COMPARING OU DATA TO BACKGROUND**

Start

Determine Background and OU Target Populations

Appropriate geographical, geological, and temporal data sets will be defined for comparison This is essentially a matching exercise so that Site (OU) data sets are comparable to background sets Consideration will be given to issues such as:

Geologic materials
Hydrostratigraphic unit
Temporal comparability
Sample size for statistical tests
Confidence in geo/hydrologic regime determination

The background data sets will be taken from the 1993 Background Geochemistry Characterization Report (EG&G, September, 1993), except for surficial soils Rock Creek surficial soil samples were used as background for OUs 1 and 2, and will be used until the FY94 surficial soil sampling data is available. Surficial soils are scheduled to be sampled in FY94 to supplement the Rock Creek data and the FY94 samples will be used subsequently as background surficial soil data The following media have defined backgrounds: groundwater (Rocky Flats Alluvium, valley fill alluvium, colluvium, weathered sandstone, and unweathered Arapahoe/Laramie formation rocks), surface water (Rock Creek and Woman Creek), seeps, stream sediments (Rock Creek and Woman Creek), seep sediments, and soils (Rocky Flats Alluvium, colluvium, surficial, weathered claystone, and weathered Arapahoe, Laramie sandstone). Site media will be cross-referenced to one or more background media.

Set DQOs

DQOs are established to define data needs for each of the RFI/RI tasks, coordinate that collection activities support those needs, and ensure the quality and quantity of resultant data. Three stages are used in the development of DQOs.

Identify Decision Types:

Identify and involve data users,
Evaluate available data,
Develop a conceptual model of the study site, and
Specify RFI/RI objectives, and anticipate the decisions necessary to achieve the objectives.

Identify Data Uses and Needs:

Identify data uses,
Identify data types,
Identify data-quality needs,
Identify data-quantity needs,
Evaluate sampling and analysis options, and
Review data precision, accuracy, representativeness, completeness, and comparability (PARCC)

Design Data Collection Program:

Assemble data-collection components, and
Develop data-collection documentation.

Data Collection and Validation

Under current IAG schedule conditions, analytical data may not be 100% "validated" when the background comparisons are made in each draft report. However, non-validated data will be used only for draft RFI/RIs Final RFI/RI reports will use only data that have undergone

27

validation Data that have been rejected will not be used The potential impacts of using non-validated data will be discussed on a case-by-case basis in the final reports

Data Presentation

A "preliminary" exploratory data appraisal will be performed to obtain a "feel" for the data This will involve techniques and identification of issues such as

Gross summary statistics
Spatial arrays
Temporal plots
Sampling strategy comparability evaluation
Affected media matrix
Hit ratios
Non-detect rates
Detection limit/quantitation limit issues
Extent of data qualifications "J", "B", etc.
Histograms/boxplots/other visuals
DQO adequacy/completeness assessment

This step will help guide the need for, and evaluate the appropriateness and applicability of further analysis, evaluate assumptions, and ascertain the impacts and limitations in light of the actual data as collected. Information generated during the exploratory data appraisal will be used in evaluating the appropriateness of the scope of the formal RFI/RI proposal. Results will be informationally discussed in a meeting with EPA, CDH, and DOE/RFO.

Several data-presentation techniques were identified by Dr Gilbert as appropriate for different conditions To perform them all for all compounds in a standard full suite is not necessary when it is clear from a preliminary review that the vast majority of data points for some compounds are entirely or almost entirely non-detects

Accordingly, we have refined the methodology as follows·

Box plots will be used when the percentage of non-detects is 50% or less.

Histograms will also be used when the percentage of non-detects is 50% or less Bars in the histogram will be shaded to indicate the percentage of detects and non-detects within each bar interval

Probability plots, ordered listings, and other graphics will be used as appropriate.

As indicated by the OU1 process, visual presentation of the data is important. Interpretable graphics will be produced to the extent that they facilitate analysis. In general, graphics will be a central feature of analysis.

BACKGROUND COMPARISON METHODOLOGY TOOL BOX APPROACH

Employing Bounding-Benchmark Comparison (Hot Measurement), Inferential Statistics, and Professional Judgement

General

The tool-box approach employs a bounding-benchmark comparison, inferential statistics, and professional judgement. This approach was forwarded in the OU1 comment-resolution process, endorsed by Dr. Gilbert, and is widely applied in the hazardous waste industry and environmental business across America It employs a "weight-of-evidence" framework wherein all three aspects are factored into the determination of what is a Site (OU) contaminant. Statisticians will be used to verify that the methods used are correct

Bounding-Benchmark Comparison ("Hot-Measurement Test" Component)

- A hot-measurement test will be performed that will compare each analyte concentration to an upper-limit value for that analyte.

- The upper-limit value will be the value at which there is a 99% probability that 99% of the background distribution will be below this value ($UTL_{99/99}$). If the $UTL_{99/99}$ cannot be calculated or reasonably estimated, then background values from technical literature and professional judgement will be used. The resulting geochemical interpretation of data will be subject to Agency review and approval.

- The $UTL_{99/99}$ is required instead of a toxicity-based value because a single list of potential contaminants must be used by many disciplines (Human Health, Ecological, Regulatory, etc ,) to ensure consistency across the RFI/RI and CMS/FS Reports The subjective nature of what is "hot", as well as toxicity and ARAR considerations, will be dealt with by the specialists who determine COC's specific to their discipline.

- In addition to ensuring that high concentrations do not get overlooked, the $UTL_{99/99}$ is an important tool for identifying locations of suspected elevated concentration in the "nature and extent" section

29

Background Comparison Using Inferential Statistical Methods

Based on Dr. Gilbert's work, the following inferential statistical tests will be used to compare background data sets to data sets compiled at the Operable Units (OUs). These data sets will be compiled and compared by analyte, and by the correct background data set (i e , colluvium, alluvium, alluvium + colluvium, surface soils, etc [See Determine Background and OU Target Populations])

It should be noted that Dr. Gilbert's recommendations establish a framework that emphasizes using the most appropriate test available. Thus professional judgement will be necessary both in application of inferential tests, as well as their interpretation Additionally, within the framework of a battery of tests drawn from a "tool box" of methods, it is requested that EPA and CDH remain open to consultation on the use of other tests as appropriate

The results of all tests (hot-measurement, inferential) will then be evaluated in light of professional judgement. This process is depicted on Figure 3, **BACKGROUND COMPARISONS METHODOLOGY**

If hot-measurement or inferential statistical tests show that the concentration of a given analyte in the OU data set is not greater than the concentration in the background data set, and if considerations in the professional-judgement arena do not override, then the analyte is considered not to be a contaminant

If either the hot-measurement test or at least one inferential statistical test shows that the concentration of a given analyte in the OU data set may be greater than the concentration in the background data set, then professional judgement (using temporal and spatial analysis, as well as pattern-recognition concepts) is again applied to see if the analyte concentrations in the two data sets are actually different.

After the hot-measurement test and prior to the use of inferential statistical testing, the issue of non-detects must be dealt with for all tests except the Gehan test, which can be applied with non-detects present. For all other tests, non-detects should be replaced with a value of 0 5 times the applicable reported detection limit, following EPA guidance (Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Addendum to Interim Final Guidance, July 1992), but realizing the performance of simple substitution decreases with an increasing proportion of non-detects

The handling of non-detects, and the presence of multiple detection limits in the RFEDS data base, requires the use of good professional judgement along with the general guidance offered here. The use of graphical displays of data will assist in the handling of high-value non-detects.

Detection limits will be discussed in the RI report

Gehan Test or Nonparametric ANOVA Test

- The Gehan test is a nonparametric test and can be used when multiple detection limits are present. The Gehan test will be applied without replacing non-detects. These are the principal favorable attributes of the Gehan test

- Standard nonparametric ANOVA tests (Wilcoxon Rank Sum and Kruskal-Wallis) are widely used in environmental assessment, and are discussed in EPA guidance (Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Addendum to Interim Final Guidance, July 1992) These tests require replacement of non-detect values, either by simple substitution or maximum-likelihood methods

- For the Gehan or nonparametric ANOVA test, a p-value will be generated and p-values that are equal to or less than 0 05 will normally be considered indicative of a significant difference from background Statements of the test and null hypotheses will be given, in both statistical and narrative terms.

Quantile Test

- The quantile test is also a nonparametric test and can be considered as a rapid screening test

- Due to limitations in the quantile test, the test will only be used if the largest 20% of the combined background and site data are detects.

- A p-value will be generated and p-values that are equal to or less than 0.05 will indicate a significant difference from background. Statements of the test and null hypotheses will be given, in both statistical and narrative terms.

Slippage Test

- The slippage test is a nonparametric test and can be considered as a rapid screening test.

- Due to limitations in the slippage test, the test will possibly not be used if the largest background value is a non-detect If the largest background value is a non-detect, then professional judgement will be applied to determine whether or not the slippage test is applicable. For example, if the second largest background value is a detect and is similar in value to the largest background value, it could be used in place of the largest value (although the replacement must be taken into account when interpreting the test results).

- A p-value will be generated and p-values that are equal to or less than 0.05 will indicate a significant difference from background. Statements of the test and null hypotheses will be given, in both statistical and narrative terms.

T-Test

- The t-test is a parametric test and is very commonly used when testing the difference between means of two data sets

- Due to limitations in the t-test, the test will be applied in cases where both background and OU data are normally distributed and contain at least 20 data points, and less than 20% of the background and OU data are classified as non-detects

- A p-value will be generated and p-values that are equal to or less than 0.05 will indicate a significant difference from background Statements of the test and null hypotheses will be given, in both statistical and narrative terms

Professional Judgement

The following general guidelines will be used individually and collectively, in conjunction with the above comparison and statistical "tools" to ascertain if a reported analytical detection(s) constitutes contamination at the OU When professional judgement is applied, documented and defensible evidence will be furnished, and DOE will bear the "burden of proof".

- **Spatial distribution** of analytes above background are or are not indicative of contamination due to waste-related activities at the OU Spatial plots, interpreted in a source-to-receptor conceptual model, in addition to compound-specific mobility considerations, generally assist in interpretation of inconclusive results.

- **Temporal distribution** of analyte concentrations at a station indicates the "high" value(s) is(are) outlier(s). Time-series plots at wells or surface-water locations can generally be used to link apparently insignificant **outlier** reports to seasonal or hydrological phenomena, and vice versa

- Other associated analytes are determined not to be contaminants in the sample or at the station Then this may be added to cumulative evidence ("burden of proof") that the analyte in question is not a potential contaminant of concern. **Pattern-recognition concepts** are useful in identifying anomalies as well as confirming "fingerprint" associations.