ndep
NEVADA DIVISION OF ENVIRONMENTAL PROTECTION
protecting the future for generations

STATE OF NEVADA

Department of Conservation & Natural Resources

DIVISION OF ENVIRONMENTAL PROTECTION

Jim Gibbons, Governor
Allen Biaggi, Director
Leo M. Drozdoff, P.E., Administrator

March 23, 2007

Ms. Susan Crowley
Tronox LLC
PO Box 55
Henderson, Nevada 89009

Re: **Tronox LLC (TRX)**
NDEP Facility ID #H-000539
Nevada Division of Environmental Protection Response to:
Upgradient Investigation Results
dated October 30, 2006

Dear Ms. Crowley,

The NDEP has received and reviewed Tronox's report identified above and provides comments in Attachment A. Once TRX has reviewed these comments it may be useful to have the NDEP's technical team discuss these matters with the TRX technical team. Please advise when a revised report can be expected.

If there are any questions please do not hesitate to contact me.

Sincerely,

Brian A. Rakvica, P.E.
Supervisor, Bureau of Corrective Actions
Special Projects Branch
NDEP-Las Vegas Office

CC: Jim Najima, NDEP, BCA, Carson City
Shannon Harbour, NDEP, BCA, Las Vegas
Todd Croft, NDEP, BCA, Las Vegas
Barry Conaty, Akin, Gump, Strauss, Hauer & Feld, L.L.P., 1333 New Hampshire Avenue, N.W., Washington, D.C. 20036
Brenda Pohlmann, City of Henderson, PO Box 95050, Henderson, NV 89009
Mitch Kaplan, U.S. Environmental Protection Agency, Region 9, mail code: WST-5, 75 Hawthorne Street, San Francisco, CA 94105-3901
Rob Mrowka, Clark County Comprehensive Planning, PO Box 551741, Las Vegas, NV, 89155-1741
Ranajit Sahu, BRC, 311 North Story Place, Alhambra, CA 91801
Craig Wilkinson, TIMET, PO Box 2128, Henderson, Nevada, 89009-7003
Kirk Stowers, Broadbent & Associates, 8 West Pacific Avenue, Henderson, Nevada 89015
George Crouse, Syngenta Crop Protection, Inc., 410 Swing Road, Greensboro, NC 27409
Nick Pogoncheff, PES Environmental, 1682 Novato Blvd., Suite100, Novato, CA 94947
Lee Erickson, Stauffer Management Company, P.O. Box 18890, Golden, Co 80402
Chris Sylvia, Pioneer Americas LLC, PO Box 86, Henderson, Nevada 89009
Paul Sundberg, Montrose Chemical Corporation, 3846 Estate Drive, Stockton, California 95209
Joe Kelly, Montrose Chemical Corporation of CA, 600 Ericksen Avenue NE, Suite 380, Bainbridge Island, WA 98110
Paul Black, Neptune and Company, Inc., 8550 West 14th Street, Suite 100, Lakewood, CO 80215
Paul Duffy, Neptune and Company, Inc., 8550 West 14th Street, Suite 100, Lakewood, CO 80215

Attachment A

1. General comment, the NDEP provides the following general comments:
   a. There is inconsistency in the report with respect to the subject-verb agreement for the usage of the word "data". In some places it is treated as singular and in others it is treated (correctly) as a plural.
   b. When statistical tests are performed, it is preferable to present the *p*-values that correspond to the test as opposed to a binary indicator of whether the null hypothesis was rejected or not. Because the *p*-value quantifies the weight of evidence against the null hypothesis, the actual value is useful when hypothesis tests are used as part of the decision-making process, as opposed to the sole determinant of the decision-making process itself.
   c. Regarding data usability, it would be helpful if TRX followed the recent example from the BRC Borrow Pit Human Health Risk Assessment for the revised version of the Upgradient Report. Please note that the revised version of the BRC Borrow Pit Human Health Risk Assessment has not been published as of the date of this letter. In addition, the NDEP would be happy to review this issue with TRX. The TRX data usability is currently incomplete.
   d. The evaluation of Data Quality Indicators is also incomplete. In particular, comparability and representativeness are insufficiently addressed.
   e. Too much reliance is placed on statistical test results, and not enough on the weight of evidence. Summary statistics and exploratory data analysis are presented, but the statistical test results dominate conclusions. They should be considered in light of the plots and summary statistics, so that informed decisions are made. This approach might shed some light on why some of the statistical results are significant and others are not. Exploration and interpretation are key, and cannot be replaced by a flowchart approach to performing statistics in a vacuum. The data can tell a story; the data analysis should expose that story. In general, this is a case where it would be helpful if some more analysis and interpretation was given. Why do some of these tests fail? Which boreholes cause the failure? Is it because they have relatively high or low concentrations? Why are depth differences seen when geological differences are not? Why are depth and geologic differences both seen for some chemicals. It is important to use the data to understand what is going on, and not simply report statistical analysis results. It is not enough to simply state that statistical tests fail or do not fail. This is a general comment that applies to all of the analyses reported.
   f. In addition, the pieces should be used to build a picture of what is happening and then there should be a report on the big picture as well. However, the presentation of results is at the level of each chemical, without building a case for what these results mean collectively. For example, there are differences between the TRX and the City of Henderson (COH) and BMI/TIMET background. This would imply that

the background distributions are different, or that there are releases impacting the site. If the latter, then it is probably inappropriate to combine data for any of the chemicals considered. This analysis is at a detail level that does not help understand what is going on at the site. The bigger picture needs to be pulled together from the pieces.

2. Table of Contents, the page numbers in the table of contents appear to be incorrect.
3. Acronyms, page iv, ANOVA typically refers to general "analysis of variance" models and not just the "one-way analysis of variance" as stated on page vi.
4. Executive Summary, the NDEP has the following comments:
   a. Page ES-1, second paragraph, first sentence states, "The upgradient investigation successfully achieved the objective of gathering sufficient soil and groundwater chemistry data to characterize the local upgradient geochemistry of the sediments in the different upgradient formations as well as to characterize the groundwater that moves through the formations." Some description of sufficiency should be presented here.
   b. Page ES-1, 1st para after the bullets, last sentence. The sentence implies the existence of groundwater background data. The sentence should be revised to make it clear that background comparisons with the COH and BMI data are only applicable to soil data. The groundwater data have not been compared to other background data.
   c. Page ES-1, 2nd paragraph after the bullets, 2nd sentence, it is not clear why this RPD objective was used. This has no statistical basis for determining the importance of differences that are observed. See other comments below on the comparison of micro-purge and bailer results.
   d. Page ES-2, TRX states "Statistical comparisons between the Tronox and COH data sets indicate that all species, except arsenic and iron, represent different populations and should not be combined for subsequent analyses." Please note that the NDEP does not necessarily concur and believes that this issue should be discussed amongst statistical personnel.
   e. Page ES-2, TRX goes on to state "This is not surprising because the COH data were collected from alluvial materials approximately 2.4 to 3.4 miles to the east of the upgradient samples." Please note that the 2.4 to 3.4 mile distance has little to do with the comparability of these samples. This issue should be discussed in terms of geochemical similarities.
   f. Page ES-2, TRX goes on to discuss the BRC/TIMET data set in a similar manner as above. Again, the NDEP believes that this issue warrants further discussion between statistical personnel.
   g. Page ES-2, 1st full paragraph, 3rd sentence. It is not clear how samples were qualified based on "representativeness". This is a qualitative issue that refers to how the samples collected represent the populations they are meant to represent. Some clarification (or deletion) would help.
   h. Page ES-2, second full paragraph, second sentence states, "The upgradient data for metals and perchlorate in soil samples were statistically compared boring to boring, depth-to-depth (20 ft or less vs. 30 ft or more), and alluvium to Muddy Creek formations." It is not clear what this means.

Perhaps the sentence could be broken into bullets that describes each set of comparisons.

i. Page ES-2, 3rd full paragraph. This and the next paragraph indicate that some of the populations are different. It is important to know more about what this means. Are the TRX concentrations greater than or less than the background concentrations in these cases? Are the differences large or small (statistical but not practical perhaps)? This gets at the general concern that too much reliance is placed on statistical test results, and that more attention should be paid to interpreting the data from summary statistics, plots and test results (including professional judgement).

j. Page ES-2, 5th full paragraph. Similar concerns about the level of interpretation provided for the statistical results that have been presented. Its also not clear if the goal here is to merge datasets, or simply to note whether the TRX and background datasets are similar or not. The background data set is quite rich at this point, so inclusion of new data in the background dataset may not be needed. In addition, since several metals and radionculides do not exhibit site concentrations that are similar to the background data, this begs the question of the reasonableness of combining any of those data. The goal instead should be comparison of the TRX data with the background data, not with a view to combination of the data for some chemicals.

k. Page ES-3, 1st full paragraph. Perchlorate is detected again below 50 feet. It would be helpful if some indication of concentrations were provided.

l. Page ES-3, 1st full paragraph. A depth is not provided for the term "shallow groundwater". It would be helpful to know the depth of the shallow groundwater here.

m. Page ES-3, 2nd full paragraph. In the context of the Executive Summary, it is not clear why a paragraph is devoted to perchlorate. Some explanation is needed for why perchlorate is called out when this is not the case for any other chemicals (except Cr).

n. Page ES-3, 2nd full paragraph. In the context of the Executive Summary, it is not clear why a paragraph is devoted to Cr. Some explanation is needed for why Cr is called out when this is not the case for any other chemicals.

5. Section 1.2.3, page 1-4, TRX states "At the request of the NDEP, soil from one boring (M-120) was analyzed for the full list of SRCs." Please revise this statement as this was never requested by the NDEP. If TRX believes that the NDEP is in error, please provide the documentation to support the above statement.
6. Section 1.2.3, page 1-4, bullets at top of page. It might be helpful to present these items on a Figure.
7. Section 1.2.3, page 1-4, last paragraph of Section 1.2.3. The borings are shown on Figure 1-2 rather than Figure 1-1.
8. Section 2.5.2, page 2-4, please note that the background summary report is currently being revised in response to NDEP comments.
9. Section 2.5.3, page 2-5, TRX refers to the NDEP's consultant as "Neptune Company". Please note that the proper company name is "Neptune and Company, Inc.".

10. Section 2.5.3, 2nd paragraph, suggest rewording the back end of sentence that states "however, the elimination of these rejected data did not adversely affect the data set statistics used in this study. It is not clear what "data set statistics" means. Perhaps the term "statistical analyses" or "data analyses" would be more appropriate.
11. Section 2.5.3, 4th paragraph, last sentence. Suggest changing "comparable" to "similar" but in the context of the distributions of the concentrations. One problem with the term comparable here is that EPA uses that term for a different purpose as one of its Data Quality Indicators.
12. Section 2.5.3, 4th paragraph, 1st sentence. Sentence does not make sense as written. It includes a clause that background data for the River range were collected because the northern McCullough range is the primary source of material… Suggest rewriting the sentence.
13. Section 2.5.3, 5th paragraph, 1st sentence. Replace test with tests at the end of the sentence.
14. Section 2.5.3, page 2-6, 1st paragraph, last sentence. It is not clear that this sentence makes sense. It is not clear what is meant by the "BRC/TIMET data set incorporates the variability of the COH data set". Perhaps this should be explained in terms of the range of the data, but variability usually means variance or standard deviation, in which case the sentence does not make sense. Some clarification is needed.
15. Section 3. It appears that the data usability step has been missed. Data validation has been performed, data evaluation has been performed, but the intermediate step as part of EPA's quality system has not been performed. See also general comment above.
16. Section 3.1, page 3-1, TRX states "The boreholes were backfilled with the unused core material". Please note that this practice is forbidden by the Nevada Division of Water Resources and should not be repeated in the future. Please note that this comment applies to similar instances discussed in other sections of the report.
17. Section 3.1, page 3-1, TRX states that a Photovac PID was used. Please discuss the bulb that was used in this PID and how this bulb relates to the ionization potential of the chemicals that were being investigated.
18. Section 3.5, page 3-5, please clarify if the wells were sampled with the bailer or micro-purge technique first. Also, please discuss the time that elapsed between each event. In addition, please discuss the amount of time that elapsed between the installation of the micro-purge well and the sampling event.
19. Section 3.12, NDEP has the following comments:
    a. Page 3-10, second to last paragraph states, "When more than two sets of data were compared, such as when the concentrations of more than two soil borings were compared, the ANOVA and the Kruskal-Wallis tests were applied." It isn't clear to NDEP that this comparison makes sense. Is this approach looking for differences between boreholes? If so, some further explanation of why this is potentially useful is needed. Is the intent to search for spatial differences in the data, so it is basically an effort at exploratory data analysis. In addition, a downside of running as many tests as have been run on the same data is that the error rate being used of 0.05 is no longer supportable.
    b. Page 3-10. The Gehan test is a generalization of the Wilcoxon Rank Sum test. That is, if there are no censored data (non-detects) then they give

exactly the same results. All that the Gehan test does that is different is provide a different ranking system for the data when non-detects are involved. Otherwise the statistical tests (Gehan and WRS) are the same. This issue seems to be missed in the presentation and in the report.

c. Page 3-10. The value of running a t-test on log-transformed data is not totally clear. Log-transformations essentially smooth the data, especially lessening the effect of higher values. Consequently, running a test that says that the mean of the logs are similar (or not) is not conceptually appealing. EPA, in its Data Quality Assessment guidance (2006) does not require testing on transformed data, but instead suggests using non-parametric tests when the normality assumptions are sufficiently violated. We would prefer that TRX performs t-tests on the untransformed data, and the WRS test (along with the Quantile and Slippage tests – see below), when comparing two sets of data, especially when one set is meant to be background. This set of tests has been long approved by EPA, and are customarily run when comparison is needed between two sets of environmental data, especially when one of the sets is a background or reference set.

d. Page 3-10, last paragraph. The NDEP does not concur with the reasons given for not running the Quantile and Slippage tests. The objectives of the statistical analysis are, in general, to determine of different sets of data (distributions of concentrations) are similar. The reason that Gilbert introduced the Quantile and Slippage tests for environmental data was precisely because it is not unusual to see differences in the tails of such distributions, when the centers are similar. Background comparisons, among other comparisons, have been performed here, and use of these tail tests is relevant and should not be dismissed without some better justification.

20. Section 3.13.1, pages 3-11 through 3-12, it is not clear to the NDEP why TRX has included an extended discussion of the data validation process in this section. NDEP and TRX have mutually agreed to a process and this should not be repeated in the revised report. This process should be summarized via a reference to the documentation between NDEP and TRX.

21. Section 3.13.2, the NDEP has the following comments:

a. Page 3-12, Section 3.13.2, 2nd bullet. It is noted that only a small number of radionuclide analyses were performed. Is this regarded as a data gap? Or, do more such data need to be collected to support hypothetical DQOs or data needs and requirements? We also note that the last sentence states, "(the comparisons for radionuclides was limited because only a small number of radionuclide analyses were conducted below the Quaternary Alluvium)." The word "was" should be changed to "were".

b. Page 3-12, Section 3.13.2, 3rd bullet. Background comparisons will be performed, but it is not clear that there is justification in combining TRX and background data sets. See earlier comments. It would be up to NDEP to decide if the background dataset should be augmented, but the arguments provided are not sufficient to justify this as a goal or objective.

Background comparisons can be performed, but the purpose should be to determine if the TRX data are similar to background.

c. Page 3-12, paragraph after bullets. It is questionable that averaging field duplicates is standard statistical procedure. Field duplicates for soil samples often should be represented as separate samples, depending perhaps on the nature of the contamination. Most metals are sufficiently particulate that field duplicates serve very little purpose for QA because they do not account for small scale variability. If the duplicates are splits (splits of a homogenized sample), then there is some QA value in their collection. A further problem is that averaging violates some basic statistical assumptions. We agree that averaging is done, but and that the assumptions violation (of independent and identically distributed assumptions) is ignored. The preference these days is to treat them as separate samples unless there is any reason not to (e.g., because they are splits). Otherwise averaging is accepted. Other options include using the first sample because the second one was collected for a different reason. From the perspective of classical statistics this is also justifiable. There is an example in EPA's Data Quality Assessment guidance (G-9, 2006) that addresses this issue, and treats the field duplicates as separate samples. Also, when it should be stated how the detection status is determined for duplicates when one of the duplicates is detected and the other is not (e.g. Sb:sample id M117-20, perchlorate M118-20)

d. Page 3-12, 2nd last paragraph. The boxplots for the "all results" for each chemical are not particularly useful. There might be better choices for showing distributions like this, such as histograms, or density estimates, but the main purpose of this data analysis is comparison, for which the side-by-side boxplots are helpful.

e. Page 3-13, second paragraph, middle. It is stated that, for reasons given, the "average arsenic concentration .... is an approximation of the true mean". This is not a correct statistical statement. Despite the fact that many statisticians do not believe in the concept of a "true mean", the average is not an approximation, it is an estimate of the "true mean".

f. Page 3-13, second paragraph, last sentence states, "Instead, statistical tests can be applied to determine with reasonable confidence if the measured concentrations came from two separate formations, even if the mean arsenic concentrations are the same or similar." The phrase "are the same or" should be omitted. If the measured concentrations from two formations are the same than there can be no statistical difference between the two.

g. Page 3-13, 3rd paragraph, third sentence states, "An appropriate statistical test could be conducted to determine the probability that the null hypothesis is true." This is technically incorrect. Statistical hypothesis tests do not compute the probability that the null hypothesis is true. Hypothesis tests are performed to determine the probability of observing a result (this result is based on the statistic of interest, and the way the data are summarized with respect to the statistic of interest) outside of the expected range of results that would be obtained when assuming that the null hypothesis is true. Basically, we assume the null hypothesis is true and then see how incongruous the data are with respect to that assumption. Perhaps the following statement could be used as a replacement: "An appropriate statistical test could be conducted to determine whether the null hypothesis should be rejected."

h. Page 3-13, fourth paragraph, last sentence states, "In contrast, nonparametric tests can be applied to any dataset regardless of the distributions." There are some distributional requirements for some non-parametric tests. For example, the Wilcoxon class of tests does technically require that the distribution be symmetric about a median. In general, non-parametric tests do not require that the distribution follow a form that can be parametrized (e.g. normal, gamma, etc).

i. Page 3-13, last paragraph, first sentence states, "If both subsets of data were assumed to follow normal distributions, the parametric F-test was conducted to evaluate if the standard deviations are equal." The F-test is performed using the variance and tests for the equality of variances. Even though the standard deviation is a function of the variance, since the test is performed on the variance, the results of the test should be interpreted in terms of the variance. An analog is that equality of the means does not imply equality of the logarithm of the means. This correction should be made in subsequent sentences as well. Additionally, it isn't clear if this test was one-sided or two-sided. This should be stated.

j. Page 3-14, first paragraph, first sentence states, "Differences among borings were evaluated using a parametric ANOVA to test the null hypothesis that the mean concentrations from all of the borings are the same and using a non-parametric Kruskal-Wallis test to test the null hypothesis that the median concentrations from all of the borings are the same." It isn't clear that this is an appropriate use of the ANOVA model. If a regular ANOVA model is run (i.e. fixed effects) then the interpretation is valid for only those borings where samples were taken. If, however, a random effects model were run, then this approach would allow for inferences among the collection of all possible boreholes.

22. Section 4.2 subsections. Please explain why comparisons have been performed with PRGs for some of the suites of chemicals and not for others.
23. Section 4.2.6, page 4-3, "Uranium (natural)" should be changed to "Uranium (elemental)."
24. Section 4.2.6, page 4-4, the summary of the radionuclide analysis presented here is fine. However, no backup is provided. These results need to be justified with the calculations that were performed. The calculations should involve some statistical analysis to demonstrate the similarities that are reported.
25. Section 4.3, page 4-5, first sentence. It is not clear that the data can lead to a conclusion about which approach leads to more representative samples. The data can lead to a conclusion that the two methods yield different results. Then a conclusion can perhaps be drawn that the micro-purge method produces more representative data, but only because there is a difference and it is believed that the micro-purge

approach is likely to give better data. That is the conclusion is based on what is expected, and then supported by the data, and not purely on the statistical evaluation. The statistics can only indicate if there is a difference.

26. Section 4.3, page 4-6, it is not clear why RPD was used for this comparison. This limits the comparison to a pair of data points at a time, does not adequately account for the direction of the differences, and the RPD provides no statistical basis for drawing conclusions. It is more appropriate, statistically, to perform a paired *t*-test (or non-parametric analog) on the paired data.
27. Section 4.3, page 4-6, paragraph in middle of page. It is stated that: "An RPD greater than 30% represents a statistically significant difference in duplicate water samples". This statement is not correct. There is no statistical significance associated with the RPD measure.
28. Section 4.5, general comment, please explain what it means that the intent is to examine potential issues related to matrix interferences? How is this done? What statistical methods are used? Is it based purely on chemistry data validation? These samples are hoped to be close to background, hence relatively unimpacted, so what is expected here? It is not clear how analysis of samples that probably will not have high concentrations of these chemicals will help when analyzing samples that have high concentrations of these analytes.
29. Section 4.5.1, the NDEP has the following comments:
    a. Page 4-7, 1st paragraph, 3rd sentence. This sentence requires some cleanup. Otherwise it seems as though silica was measured in 45 samples. Use of semi-colons to separate items might help.
    b. Page 4-7, 1st paragraph, 4th sentence. The way the sentence is worded makes it seem as though perchlorate is a metal. Perhaps the sentence can be reworded.
    c. Page 4-7, second paragraph, first sentence states "Box and whisker plots of the data for each metal and for perchlorate in the soil samples are presented in Figure 4-7." The legend in figure 4-7 states that the whiskers of the boxplot extend to the minimum and maximum value. This is incorrect. The third to last sentence in this paragraph correctly states "The whiskers extend to the largest and smallest values that are not more then 1.5 times the IQR range above or below the box." The same changes need to be made to the legends in Figures 4-8 through 4-15.
    d. Page 4-7, second paragraph, last sentence. Note that the box plots as presented show the mean concentration as well.
    e. Page 4-8, first sentence states, "Box and whisker plots for metals and perchlorate grouped by boring are presented in Figure 4-8." It isn't clear how or when multiple samples were collected from within each borehole. Please clarify if these samples from multiple depths within the same borehole.
    f. Page 4-8, 1st paragraph, in looking at some of the plots, some of the ANOVA results are "unexpected". This is a case where it would be helpful if some more analysis and interpretation was given. Why do some of these tests fail? Which boreholes cause the failure? Is it because they have relatively high or low concentrations? It is not enough to simply

state that statistical tests fail or do not fail. This is a general comment that applies to all of the analyses reported.

g. Page 4-8, second paragraph, first sentence states, "Box and whisker plots grouped by sample depth are presented in Figure 4-9." Why is only a subset of the analytes presented in Figure 4-9?
h. Page 4-8, second paragraph, last sentence states, "Based on the apparent differences in concentrations between these two depths, statistical tests were conducted to compare subsets of the data in these two depth ranges." There should be a reference to the table where the results of these statistical tests are presented. Additionally, is there a physical reason that these differences between data greater than 20ft and less than 20ft exist? It isn't clear that dividing the data based on observed differences and then running statistical tests to quantify these differences makes sense in the absence of a physical reason for differences that can be incorporated into the conceptual model.
i. Page 4-8, 3rd paragraph, if only 3 samples were collected from the fine-grained facies, did TRX also consider removing them from the analysis? Please consider if it would make any practical difference in the statistical results.
j. Page 4-8, bullets. This separation is curious. The separation by depth needs to compared to the separation by geology. That is, perhaps when both distinctions occur they are for the same basic reason. This should be investigated further in an attempt to simplify this process of separating data sets. When there are statistical differences in one case and not the other, is it because the difference is marginal statistically. Presumably the data are being split similarly for these 2 cases (depth and geology), at least there must be overlap, in which case it is worth exploring the data further to understand what the results of the statistical analyses are trying to say.
k. Page 4-8, fifth paragraph, first sentence states, "Differences were statistically significant by depth range but not by geological formation for two chemicals: tungsten, vanadium, and perchlorate." Tungsten should be removed.
l. Page 4-8, last bullet on the page states, "If differences were statistically significant by both depth range and geological formation, preference was given to the categorization (i.e., by depth range or by geological formation) that resulted in subsets of the data that followed either normal or lognormal distributions. This selection was made to provide subsets of the data that could be used in parametric statistical tests for future comparisons. If both categorizations led to subsets that followed normal or lognormal distributions, the data were categorized by geological formation. Similarly, if neither categorization led to subsets that followed normal or lognormal distributions, the data were also categorized by geological formation." The decision process for partitioning should account for a conceptual understanding of the site as opposed to convenience for statistical testing. For example, differences as a function of both depth and geology are not surprising since geology is a function of

depth. The existence of normal distributions for both subsets of data defined as a function of geology, provides some evidence that the differences are due to geology as opposed to anthropogenic contamination that is diluting as a function of depth and hydrogeology (e.g. perchlorate). However, the existence of normal distributions for both subsets of the arsenic and potassium data defined as a function of depth suggests that something is missing from the conceptual model. For example, is 20 ft the vertical extent of groundwater rise during anomalous precipitation events? Additionally, it seems odd that so many analytes have lognormal distributions for both subsets of the data defined as a function of depth (e.g. barium, chromium cobalt, magnesium, uranium, and vanadium).

m. Page 4-9, "Upgradient Data vs. Background Data" section. There should be some brief review of the relevant aspects of the COH and BRC/TIMET datasets here. Specifically, what are the depths for the COH and BRC/TIMET datasets and why is it meaningful to compare the TRX data to the COH and BRC/TIMET datasets?
n. Page 4-9, Section 4.5.1. Given the results that for many chemicals there are statistical differences between geologies or depths, and between TRX data and background, it is more reasonable, in a bigger picture sense, to conclude that TRX and background data sets should not be merged. It would be very difficult to justify merging for some chemicals and not others, when the differences that exist can be due to releases as well as to geology differences. If there are any releases in this area, then background conditions as a whole do not exist, and combination of TRX and background data sets may not make sense.
o. Please discuss if TRX considered comparing the upgradient data only to the McCullough Mountains data set from the BRC/TIMET/COH data set.

30. Section 4.5.2, page 4-9, earlier it was indicated that the radionuclides are in secular equilibrium. However, in this section some radionuclides are considered greater than background and others are not. Are there any further observations that can be made to clarify the interplay between the background comparisons and secular equilibrium?
31. Section 5.0, the NDEP has the following comments:
    a. Page 5-1, Data Validation section, reference is made to data quality indicators, however, it is not clear how the issues of representativeness and comparability were dealt with or if there is any effect from them on the results and conclusions.
    b. Page 5-1, statistical evaluation section, last sentence states, "For this reason, the data for these 15 metals and perchlorate from the specific geologic formation, alluvium, or Muddy Creek Formation, or from specific ranges of depth, 20 ft or less or 30 ft or more, should be used separately for future comparisons with downgradient data." Based on the previous two sentences, this statement does not make sense. Specifically, which 15 metals are referenced? Additionally, it is not clear how the results of the differences among borehole analysis are useful in a decision-making context.

c. Page 5-1, section “Statistical comparison with Off-Site Data Sets”, second sentence states, “Statistical comparisons between the Tronox and COH data sets indicate that all species, except arsenic and iron and selenium represent different populations and should not be combined for subsequent analyses.” This conclusion for Selenium needs to be supported by additional interpretation of results found on Page 4-9, section Upgradient Data vs. Background Data, paragraph 2. Note again, given this analysis a more reasonable conclusion is that the TRX and background datasets should not be combined.

d. Page 5-1, section “Statistical comparison with Off-Site Data Sets”, last sentence states, “Because arsenic, iron and selenium concentrations did not exhibit statistically significant differences in their mean or median concentrations or standard deviations, those parameters, for the samples collected at depths of 20 ft or less, from the COH and Tronox datasets can be combined for subsequent analysis.” The results for differences in standard deviation for subsets of the data have not been presented or discussed in the text.

e. Page 5-2, background comparisons in general. Comparability is a very important issue for comparing two different data sets. There should be some discussion of this issue.

f. Page 5-2, first sentence states, “Statistical comparisons between the Tronox and BRC/TIMET data sets indicate that all species, except calcium and lead, represent different populations and should not be combined for subsequent analyses.” The reasoning for not combining any of the analytes except calcium or lead needs to be better explained either here or on Page 4-9, section “Upgradient Data vs. Background Data”, third paragraph. Specifically, Page 4-9, section “Upgradient Data vs. Background Data”, third paragraph, first sentence states ” Differences between the means or medians of Tronox and BRC/TIMET data are not statistically significant for 11 of the 27 chemicals that were measured in both studies.” However, 9 of the 11 chemicals (excluding lead and calcium) are not discussed.

g. Page 5-2, second paragraph, last sentence states, “Statistical comparisons between the Tronox and BRC/TIMET data sets indicated that data for thorium 230 and uranium 234 could probably be combined for subsequent analysis.” This conclusion is made based on results presented in Page 4-10, section “Upgradient Data vs. Background Data” second paragraph, although this paragraph does not explicitly state which datasets are being compared to obtain these results. Previous more general comments about combining datasets apply, again.

h. Page 5-2, section “Groundwater Sampling Comparison”, second paragraph, first sentence states, “In general, the less soluble constituents appear to be affected more than the highly soluble constituents.“ It should be mentioned in this statement that differences in measured concentrations between methods is a function of solubility.

i. Page 5-2, Evaluation for matrix effects section. Again, it is not clear exactly what the purpose is of this evaluation.

j. Page 5-2, Groundwater Sampling Comparison section, "Perchlorate", states "Below a depth of 20 ft bgs, perchlorate was not detected in soil samples until 50 ft bgs, which suggests that the perchlorate at this depth in soil is not related to vertical downward migration of shallow sources but is related to the perchlorate in the groundwater." Is it possible that the decreased concentrations observed above ground water but below 20 ft. are a consequence of fluctuations in the water table that "wash" the perchlorate out of the soil and into ground water? Also, the text indicates that perchlorate is present upgradient. Isn't there also an onsite source? Some clarification would help.

32. Figure 3-1. The diagram provides a flow path for statistical analysis steps. The first problem with this type of approach is that it takes professional judgement out of the decision making process. Exploratory data analysis and statistical test results are disjointed, which is also evident in the main report. In addition, many statistical tests are performed on the same subsets of data, in which case a different *p*-value should be used if an omnibus *p*-value of 0.05 is desired. Simplification is possible by not performing log –transforms ,which can only lead to conclusions in the log-space, so they are not very useful. The final conclusions are based on the test statistic results with a straight comparison to a *p*-value of 0.05. Apart from probably being the wrong *p*-value to use in the context of family-wise error rates, a straight comparison without revisiting the data implies a lack of interpretation of the entire statistical package that is offered. This is evident in the main report. Much more needs to be made of all of the statistical tools and analyses.
33. Figure 4-7 by itself is not very useful. Other ways of displaying single distributions could be used, such as histograms and density estimates, but the basic issue remains. Single plots of the combined TRONOX data are not very helpful.
34. Table 4-4. For Well IDs H-11 and M-117, the detection limit is 16 µg/L, which is four times greater than the detection limit for TR-07 and TR-09. Additionally, since the USEPA PRG is equal to 4 µg/L, the utility of these samples may be limited.
35. Appendix E, the NDEP has the following comments:
    a. General comment, the groundwater radionuclide data is not in secular equilibrium. Please discuss this matter in the main body of the report.
    b. The NDEP's review of this Appendix included a supplemental deliverable that was provided by TRX. Please include this information in the finalized report.
    c. Table E-6 contains a column labeled "Results." However, these are not actually the sample concentrations but the reporting limits in most cases. The Table should clarify this discrepancy.
    d. Section 3.3, page 7. The report states, "No data from the SW-846 601B analyses …" Please revise "601B" to "6010B".
    e. Section 3.4 and General, regarding trip blanks, the report states, "No data required qualification due to trip blank contamination." However, there is confusion whether trip blanks were included with these samples. Section 3.8.1 of the main report indicates trip blanks were included in the field QA/QC. However, the data validation memos labeled "TH021voclms.rev" and "TH018voclms.rev" indicate that no trip blanks

were submitted. The data validation report should clarify if, or for which sample sets, trip blanks were included for the VOC analysis.

36. Appendix F, the NDEP has the following comments:
    a. Section 1.1, page 1-1, Item 1. Was the Gehan ranking scheme also used for the Kruskal-Wallis test when non-detects were involved?
    b. Section 1 subsections. There is a lot of redundancy in these subsections, suggesting that the subsections could be reorganized to reduce repetition.
    c. Other statistical comments have been made in the main text, but they apply equally here.
    d. Section 1.1, page 1-1, subsection 1, sentence 1 states, "The results from an Analysis of Variance (ANOVA) to compare the mean concentrations of the chemical by soil boring and the results from a Kruskal-Wallis test to compare the median concentrations by soil boring." If a regular ANOVA model is run (i.e. fixed effects) then the interpretation is valid for only those borings where samples were taken. If, however, a random effects model were run, then this approach would allow for inferences among all possible boreholes.
    e. Section 1.1, page 1-2, number 6b, first sentence states "If both sets of data were considered to follow lognormal normal distributions, a t-test was performed on the logarithms of the data to compare the means of the logarithms of the data." First, it is not clear what it means for data to follow a "lognormal normal" distribution. Second, it is not clear that it is of interest to detect differences between the means of the logarithms of the data. Differences in the means of the logarithms of the data are not equivalent to differences in the means of the untransformed datasets.
    f. Page 2-17 appears to have a graphics error.
    g. Table F-1. The title has a typo. TRONOX is spelled TONOX.
    h. Comment 12a of the meeting minutes from 1/16/2007 states "It was noted that the TRX upgradient data showed conformance with the BRC/TIMET background data set via the box and whisker plots but not via the quantitative statistical tests." The test results appear to have been interpreted correctly. Since the tests were performed as two-sided tests, significant differences will be indicated if, for a given analyte, either the center of the distribution of the Upgradient data is greater than center of the distribution for the BRC/TIMET data or the center of the distribution of the BRC/TIMET data is greater than center of the distribution for the Upgradient data. This is a possible reason for the confusion.

37. Appendix I, the NDEP has the following comments:
    a. 1st subsection titled "Historical Groundwater Sampling". The first sentence makes a statement that is not achievable from the data analysis. The data analysis can point to a difference, but the nature of the difference can only be provided by a conceptual understanding of why it occurred. The difference cannot by itself point to a conclusion of which method is most representative.

b. Other statistical comments have been made in the main text, but they apply equally here. These pertain mostly to the need to run paired *t*-tests instead of relying on RPD.
c. Page 3 of 3. For example, arsenic is classified as a metal that did not meet the RPD standard. However, it failed in only 1 of the 6 pairs. Considering the data as a whole would lead to a different conclusion for arsenic (i.e., that, statistically, there are no differences).
d. Page 1, based on this memorandum it appears that the wells were sampled via a bailer, a micro-purge pump was installed and then the well was sampled via micro-purge techniques. The specific timing of these activties needs to be discussed. Please note that these activties would result in a large amount of agitation (and volatization) within the well. These issues should be discussed in the body of the Appendix.
e. Page 2, since TPH, VOCs, and other compounds were not detected, this study was of limited use. The volatile compounds are of particular interest when discussing bailers and micro-purge techniques. Metals and radionuclides are also of interest and the study did note significant differences in these analyses.
f. Page 3, TRX summarizes the results of the study but does not draw any significant conclusions. For example, the study does demonstrate that bailing does bias some metals and radionuclides artificially high. In addition, it appears that bailing does bias some VOCs artificially low. It would benefit TRX to utilize the micro-purge technique to produce more representative data.
g. Additional comments on the micropurge method are provided below:

Low flow purging and sampling is a method of collecting a "representative" sample using the maximum flow rate that causes minimum drawdown; thereby, minimizing the stress to the groundwater system. Mobile colloid particles ranging in size from 1 to 1,000 nm have been observed under different conditions. For a sample to be considered representative of the formation water, the sample should contain the total mobile contaminant loading that includes both the dissolved contaminants and the naturally suspended particles (Puls & Barcelona 1996; Powell & Puls 1997; Kearl et al. 1994). Using low flow purging and sampling helps prevent the entrainment of larger, not naturally mobile particles into the groundwater. Low flow purging and sampling are applicable for various contaminants and naturally occurring analytes including volatile and semi-volatile organic compounds (VOCs and SVOCs), metals, other inorganic compounds, pesticides, polychlorinated biphenyls (PCBs), other organic compounds, radionuclides, and microbiological constituents. Low flow purging and sampling are not applicable for non-aqueous-phase liquids (ASTM 2002; Yeskis & Zavala 2002; Richey 2002, FDEP 2003).

The typical range of flow rates vary from 0.1 to 0.5 L/min. Some high permeability formations may be able to use flow rates as high as 1 L/min (US EPA Region 1 1996; Powell & Puls 1997; ASTM 2002; Ritchey 2002; Kaminiski 2003). The actual flow rate and amount of drawdown that may be sustained for a particular monitoring well should be determined prior to sampling. A stabilized pumping water level should be achieved with minimal

drawdown (to minimize stress to the system) at as high a flow rate as possible (to minimize sampling time). Minimizing turbulence should also be considered when selecting a flow rate (Barcelona et al. 2005). Minimal drawdown may vary from inches for high permeability formations to several feet for low permeability formations (FDEP 2003; Barcelona et al. 2005). The flow rate should not be determined by assigning an arbitrary number for acceptable drawdown. Minimal drawdown and corresponding flow rate will be dependent upon hydrogeologic setting and well construction characteristics (Barcelona et al. 2005).
The advantages of low flow sampling are collection of groundwater samples that are representative of the mobile contaminant load, minimization of sampling artifacts, less operator variability with greater operator control, minimization of stress on formation, minimization of mixing of stagnant casing water with formation water, reduced need for filtration of samples, reduced waste generation, and higher sample consistency (NMED 2001; Puls & Barcelona 1996). The disadvantages of low flow sampling are higher initial capital costs, longer set-up time in field, additional equipment to transport, and increased training of staff (Puls & Barcelona 1996). It should be noted that the costs of obtaining representative groundwater samples may be insignificant to the costs of potential remediation decisions made based on the data collected from the samples (Yeskis & Zavala 2002).

Metals sampling should not be conducted with bailers due to increased turbidity, which may bias metals concentrations high if the samples are not filtered (Yeskis & Zavala 2002; Kaminiski 2003). However, filtering samples may bias metal concentrations low due to the filtration of naturally mobile suspended solids (Puls & Barcelona 1996; Browner, 1997). Filtering of samples has also been shown to produce inconsistent results in terms of metals mobility (Kearl et al. 1994). No filtration or sampling method exists to restore data quality of a groundwater sample after the aquifer matrix and/or sand pack has been disturbed during purging / sampling and turbidity has been artificially increased (Powell & Puls 1997). Sampling with a bailer may also bias metals concentrations by the agitation of groundwater during the insertion and removal of the bailer, causing the introduction of air into the well bore and consequently cause some metals to precipitate (Kaminiski, 2003). VOC sampling should not be conducted with the use of bailers, which may bias VOC concentrations low due to the agitation of the groundwater and the introduction of air into the groundwater within the well (NMED 2000, US EPA Region 4 2001; Yeskis & Zavala 2002; Kaminiski 2003).

Dedicated sampling pumps are recommended for low flow purging and sampling to avoid the generation of excess turbidity caused by insertion of the sampling pump thereby mixing the stagnant water in the casing above the screen with the screened interval water zone. Additionally, insertion of a portable system may cause the resuspension of solids that may have collected at the bottom of the well (US EPA Region 9 1995; Puls & Barcelona 1996; NMED 2000). Dedicated sampling pumps are also recommended to reduce the amount of waste material generated by minimizing purge volume required for stabilization of water quality indicator parameters. The time required for set-up and purging is also reduced with the dedicated systems (Puls & Barcelona 1996). Dedicated sampling pumps would not be as important in wells screened across the water table as for wells with submerged screens where stagnant water would exist above the screen interval. If dedicated sampling pumps cannot be left in-place, then the sampling pump should be slowly lowered into the screened interval to minimize mixing followed by immediate low-flow purging and sampling (Powell & Puls 1997).

Recent research has demonstrated that the entire screened interval is sampled during a low-flow purging independent of pump placement within the screened interval. Additionally, this research demonstrated that the ratio of flow yielded by higher permeability layers versus lower permeability layers is independent of pump placement within the screened interval (Varljen et al. 2006).

References

ASTM International. 2002. Standard Practice for Low-Flow Purging and Sampling for Wells and Devices Used for Ground-Water Quality Investigations. *Annual Book of ASTM Standards* D 6771-02.

Barcelona, M.J., M.D. Varljen, R.W. Puls, and D. Kaminski. 2005. Ground Water Purging and Sampling Methods: History vs. Hysteria. *Ground Water Monitoring & Remediation* 25 no. 1: 52-62.

Barcelona, M.J., J.P. Gibb, J.A. Helfrich, E.E. Garske. 1985. Practical Guide for Ground-Water Sampling. *Illinois State Water Survey.*

Browner, C.M. 1997. To Filter, or Not to Filter; That is the Question. *Letter from US EPA.*
Florida Department of Environmental Protection (FDEP). 2003. Low-Flow/Low Volume Purging and Sampling of Ground-Water Monitoring Wells Performance and Application Criteria, Version 8. *www.dep.state.fl.us*

Giles, G. and J. Story (New Jersey Department of Environmental Protection). 1997. The Low-Down on Low Flow. *Site Remediation News* 9 no. 3.

Kearl, P.M., N.E. Korte, M. Stites, and J. Baker. 1994. Field Comparison of Micropurging vs. Traditional Ground Water Sampling. *Ground Water Monitoring & Remediation* 18 no. 3: 138-190.

Maine Department of Environmental Protection (MDEP). 1996. Information Sheet Low Flow Ground Water Sampling. *Maine.gov.*

New Mexico Environment Department (NMED), Hazardous Waste Bureau. 2001. Position Paper: Use of Low-Flow and Other Non-Traditional Sampling Techniques for RCRA Compliant Groundwater Monitoring.

New Mexico Environment Department, Underground Storage Tank Bureau. 2000. New Mexico Underground Storage Tank Bureau Guidelines for Corrective Action.
OLQ Geological Services (OLQ). 1998.

Technical Memorandum Prepared for the State of Indiana: Short Review of the Micro-Purging Option for Monitoring Wells. *www.ai.org/idem/*

Parker, L.V. and T.A. Ranney. Sampling Trace-Level Organic Solutes with Polymeric Tubing, Part 2. Dynamic Studies. *Ground Water Monitoring & Remediation* 18 no. 1: 148-155.

Powell, R.M. and R.W. Puls. 1997. Hitting the Bull's-eye in Groundwater Sampling. *Pollution Engineering*

Puls, R.W. & M.J. Barcelona. 1996. Low-Flow (Minimal Drawdown) Ground-Water Sampling Procedures. *Ground Water Issue* EPA/540/S-95/504.

Ritchey, J. 2002. Low-Flow Purging and Sampling Ground Water, Evolution of Technology and Standards. *ASTM Standardization News* Apr 2002: 21-25.

United States Environmental Protection Agency Region 1 (US EPA Region 1). 1996. Low Stress (low flow) Purging and Sampling Procedure for the Collection of Ground Water Samples from Monitoring Wells, Revision 2.

United States Environmental Protection Agency Region 4 (US EPA Region 4). 2001. Environmental Investigations Standard Operating Procedures and Quality Assurance Manual. *www.epa.gov/region4/sesd/eisopqam/eisopqam.html.*

United States Environmental Protection Agency Region 9 (US EPA Region 9). 1995. Quick Reference Advisory, Use of Low-Flow Methods for Ground Water Purging and Sampling: An Overview.

Varljen, M.D., M.J. Barcelona, J. Obereiner, and D. Kaminski. 2006. Numerical Simulations to Access the Monitoring Zone Achieved during Low-Flow Purging and Sampling. *Ground Water Monitoring & Remediation* 26 no. 1: 44-52.

Yeskis, D. & B. Zavala. 2002. Ground-Water Sampling Guidelines foe Superfund and RCRA Project Managers. *Ground Water Forum Issue Paper* EPA 542-S-02-001.