APPENDIX P
Comparing Laboratory and Field Data

P-1. Introduction. Interpreting field data may arise in the SI and RI phases of a CERCLA project. The following discussion applies to comparing field data results to laboratory results.

P-1.1. As previously discussed, there is an inherent relationship among variability, the statistical decision confidence required, and the number of data points one must have to make the decision. There is a trade-off between cost and data quality (level of confidence for the decision-making). In general, cost and the level of confidence increase as the number of samples increases. In fact, a small set of very high quality individual measurements (e.g., from a fixed-laboratory analytical method) is frequently not as desirable as a large number of lower quality measurements (e.g., from a field analytical method). If rapid and inexpensive methods of sampling and analysis were available for the SI, a larger number of samples could be used to characterize the study area, reducing both cost and decision uncertainty. However, such methods with sufficient reliability are not always available.

P-1.2. There are many innovative field-based sampling and analysis techniques and technologies available to environmental scientists. Because of the ability to reproduce these sampling techniques with an acceptable level of accuracy and at relatively low cost, investigators can still make decisions with confidence based on field analyses.

P-1.3. When applying field analytical technologies to a given site, the project team often collects larger sample aliquots for a percentage of the field samples to ensure that the field methods are providing reasonably precise, accurate, and representative results. Each aliquot is thoroughly homogenized (i.e., unless VOCs are being analyzed) and split into a pair of duplicate samples; one sample is analyzed by the field method and the remaining sample of the duplicate pair is sent to a fixed laboratory for analysis. The results of the laboratory and field analyses are then compared to assess the usability of the field results.

P-1.4. Although the EPA has generally specified splitting 10% of screening samples with a fixed laboratory for confirmation analysis, this is an arbitrary criterion. Furthermore, there is little guidance on how to compare field and fixed laboratory results and the criteria for acceptable agreement. Therefore, a number of possible approaches are available and discussed here, including the following.

P-1.4.1. Relative percent difference (RPD).

P-1.4.2. Correlation analysis.

P-1.4.3. Regression analysis.

P-1.4.4. Group comparisons.

P-1.4.5. Percent decision match.

P-1.5. Project planners should be sensitive to the possible comparison methods so that sampling design is appropriate for the data collected and the decision to be made at their particular site.

P-2. Relative Percent Difference. The RPD for a duplicate pair of measurements (x_1, x_2) is the absolute value of the difference between the measurements divided by the mean of the measurements $\bar{x}$, expressed as a percentage:

$$RPD = \frac{|x_1 - x_2|}{\bar{x}} \times 100 \ .$$

P-2.1. The RPD is simple to calculate and has historically been used to compare two sets of data. The field values and the corresponding laboratory values are treated as duplicate pairs, and an RPD is calculated for each pair. It should be noted that, as it is usually used for environmental applications, the RPD is not a statistically based measure of agreement. The approach is semi-quantitative at best, and, in general, is not recommended. Acceptance limits for the RPDs tend to be arbitrarily defined and unrelated to acceptable tolerances for uncertainty (i.e., the RPD acceptance limits are not derived from statistically based data quality objectives for the project). Furthermore, the EPA has not established fixed acceptance limits for the RPDs of field duplicates, though EPA Region II has specified field duplicate acceptance limits for metals for data review.

P-2.2. The RPD limit for field duplicates is 50% for water and 100% for soils. RPD values from intra-laboratory studies are available for most SW-846 methods, but the values represent only the analytical component of the variability. As the RPD is proportional to the absolute difference, it is not useful for evaluating bias. Moreover, in terms of project decision-making, a process has not been developed to readily quantify the uncertainty associated with field results, nor has a range of acceptable RPD results been developed to determine whether field results are within decision limits.

P-3. Correlation Analysis.

P-3.1. Field data can be compared to confirmation data, typically fixed laboratory data, using correlation analysis. In this case, the data are paired and plotted on a graph, and a *Pearson's r*,[*] which is a measure of the degree of linear association between the two sets of data, is calcu-

[*] Appendices O and Q.

lated. Paired statistical tests are useful because they can be used to determine whether a screening-level method is producing data that are significantly different from a definitive method. Higher values of Pearson's *r* are preferred, as this indicates increasing similarity between the field and confirmation data. For sufficiently high values of Pearson's *r*, the field data can reliably be used as a proxy for the confirmation data. As previously stated, there are no fixed limits for comparison, but Appendix O provides some guidance for assessing correlation results in terms of values of Pearson's *r*.

P-3.2. However, there are a number of problems with using correlation analysis as a comparison tool. A principal problem is that correlation does not imply a cause-and-effect type of relationship or provide predictive capabilities. In other words, correlation analysis cannot be relied upon to show how variable *X* affects variable *Y*, or how *X* is a predictor of unknown values of *Y*. Thus, correlation analysis is intended as a statistical tool to simply show how two variables are linearly related and the strength of this relationship. An additional problem, or complexity, with correlation analysis is that the principal statistic reported in the analysis, Pearson's *r*, requires the *X* and *Y* variables to possess a *bivariate normal distribution*[*] (not only must *X* and *Y* be normal but the "joint variation" must also be normal; that is, if every possible (*x*, y) pair were available, *Y* must be normal for every fixed value $X = x$ and *X* must be normal for every fixed value $Y = y$). Finally, it is entirely possible that data sets paired in order of concentration will show linear correlation when the absolute differences between them are very large, but in some manner proportional. Thus, along with other measures, if the data give a good linear or curvilinear fit with strong correlation, this may be taken to support but not prove confirmation between results.

P-4. Regression Analysis. Field data are often compared to confirmation data, typically fixed laboratory data, using regression analysis. In this case, the data are paired and plotted on a graph and a best-fit line is created. The regression model can provide information regarding the magnitude of the difference or the functional relationship between the screening-level and definitive methods, so that screening-level data can be converted to definitive data.

P-4.1. However, functional relationships between screening-level and definitive data are often inappropriately established. Classical linear regression analysis, as presented in Appendix O, is not appropriate for this analysis because both screening-level data (the "dependent" variable) and laboratory concentrations (the "independent" variable) are measured values, and because the laboratory concentrations (the "independent" variable) has more than a negligible amount of variability. For example, the laboratory concentrations could be selected as the "independent" variable *X* to generate a regression line of the form,

$$y = b_1 x + b_0 .$$

[*] Appendix O.

P-4.2. This implies

$$x = (1/b_1)y + (-b_0/b_1).$$

P-4.3. However, the alternative selection of *Y* as the "independent" variable would produce a regression line,

$$x = b_1' \ y + b_0' \ .$$

P-4.4. Unfortunately, $b_1' \neq 1/b_1$ and $b_0' \neq (-b_0/b_1)$. In other words, the classic or ordinary least squares (OLS) line produced from *X* and *Y* measurement data depends upon whether *X* or *Y* is arbitrarily selected as the independent variable. Therefore, it would be inappropriate to generate a regression line to "convert" screening level measurements to laboratory concentrations (or vice versa).

P-4.5. In place of OLS linear regression, reduced major axis (RMA) regression is a reasonable parametric approach, while the Kendall-Theil line is a desirable non-parametric approach for establishing a linear relationship. Advantages to reduced major axis regression are the following.

P-4.5.1. While a classic (OLS) regression line of the form $y = b_1x + b_0$ minimizes the sum of the distances in the *y*-direction from the regression line to each observed point y_i, the RMA line minimizes error for both *X* and *Y* by minimizing the sum of the areas of right triangles formed by horizontal and vertical lines extending from each observation (x_i, y_i) to the best-fit straight line (Helsel and Hirsch, 1992, p. 276).

P-4.5.2. Unlike OLS regression, RMA regression produces a unique line regardless of which variable, *X* or *Y*, is used as the response or independent variable.

P-4.6. RMA regression is used to model the correct functional relationship between two variables when both variables possess comparable measurement error. It is commonly used to evaluate biological data. All of the assumptions required for OLS regression are required for RMA regression (e.g., the residuals must be normally distributed). RMA regression has also been called "line of organic correlation," "geometric mean functional regression," and "Maintenance of Variance-Extension" (Helsel and Hirsch, 2003). Reduced major axis regression should not be confused with an alternative approach referred to as "major" or "principal axis" regression. Major axis regression is often used in lieu of RMA regression as it is conceptually similar; the best fit line minimizes the sum of the squares of the perpendicular distances between the line and each plotted observation (rather than the areas of right triangles). Both reduced major axis

and major axis regression are often referred to as "model II" regression (OLS regression is "model-I" regression).

P-4.7. The slope (b_1'') and intercept (b_0'') of the RMA regression line $y = b_1''x + b_0''$ are as follows:

$$b_1'' = sign[r]\left(s_y / s_x\right)$$

$$b_0'' = \bar{y} - b_1''\bar{x}$$

where $sign[r]$ is the algebraic sign of Pearson's r; s_y and s_x are the sample standard deviations of Y and X, respectively; and $\bar{y}$ and $\bar{x}$ are the sample arithmetic averages of Y and X, respectively. Like an OLS regression line, the RMA regression line passes through the point $(\bar{x}, \bar{y})$, but (unlike an OLS regression line) the slope does not depend upon the magnitude of the regression coefficient r. Given the OLS regression lines $y = b_1 x + b_0$ and $x = b_1' y + b_0'$, an alternative expression for the major axis regression slope is:

$$b_1'' = sign[r]\sqrt{b_1 / b_1'}\ .$$

P-4.8. Thus, the slope of the RMA regression line is essentially the geometric mean of the OLS slopes b_1 and $1/b_1'$ (hence the use of the terminology "geometric mean regression"). An equivalent expression for the RMA slope is:

$$b_1'' = b_1 / r$$

Note that, because $r \leq 1$, the RMA slope will be equal to or greater than the slope of the corresponding OLS regression line.

P-4.9. Confidence limits can be calculated for the slope and intercept of the RMA regression line. The $(1 - \alpha)100\%$ confidence interval for the slope is as follows (Warton, 2005)

$$\left[b_1''\left(\sqrt{B+1} - \sqrt{B}\right),\ b_1''\left(\sqrt{B+1} + \sqrt{B}\right)\right] \tag{P-1}$$

where

$$B = \frac{F_{1-\alpha,1,n-2}\,(1 - r^2)}{n-2}$$

$F_{1-\alpha,1,n-2}$ is the critical value of the *F*-distribution with 1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator. The confidence limits for the intercept are:

$$b_0'' \pm t_{1-\alpha/2,n-2}\, s_0 \quad . \tag{P-2}$$

P-4.10. The quantity s_0 denotes the estimated standard deviation of the intercept of the OLS regression line $y = b_1 x + b_0$, which may be determined from the equation:

$$s_0 = \sqrt{\frac{s^2}{n} + \bar{x}\, s_1^2} \quad .$$

P-4.11. The quantity s^2 denotes the estimated variance of residuals of the OLS regression line $y = b_1 x + b_0$ and s_1^2 the estimated variance of slope of the OLS slope

$$s^2 = \frac{s_y^2 (n-1)\left(1-r^2\right)}{(n-2)}$$

$$s_1^2 = \frac{s^2}{s_x^2 (n-1)}$$

where

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

and

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{(n-1)} \, .$$

P-4.12. The reader is referred to software that can be used to calculate RMA regression lines as well as confidence limits for the slopes and intercepts (Bohonak, 2004), though the software does not calculate the confidence limits of the slope using Equation P-1 but using an approximation that produces a similar result:

$b_1'' \pm t_{1-\alpha/2,n-2}\, s_1$.

P-4.13. A non-parametric approach for establishing a linear relationship is the Kendall-Theil line. The line takes the form: $y = \hat{b}_1 x + \hat{b}_0$. The slope $(\hat{b}_1)$ is computed by comparing each data pair to all others in a pairwise fashion. A data set of n (x, y) pairs will result in $n(n-1)/2$ pairwise comparisons. For each of these comparisons, a slope is computed by

$$m_{ij} = \frac{\left(y_j - y_i\right)}{\left(x_j - x_i\right)} \text{ for all } i < j;\ i = 1,2,\ldots,(n-1); \text{ and } j = 2,3,\ldots,n.$$

P-4.14. Note that m_{ij} is the value of the random variable, M. The slope $(\hat{b}_1)$ and intercept $(\hat{b}_0)$ are estimated as follows:

$\hat{b}_1 = \tilde{m}$, where $\tilde{\text{m}}$ is the median of M

and

$\hat{b}_0 = \tilde{\text{y}} - \hat{b}_1 \tilde{x}$, where $\tilde{\text{y}}$ and $\tilde{\text{x}}$ are the medians of Y and X, respectively.

P-4.15. Therefore, the line passes through the point $(\tilde{x}, \tilde{\text{y}})$, analogous to the ordinary least squares regression line, which passes through the point $(\bar{x}, \bar{y})$. The Kendall-Theil line is closely related to the Kendall's τ (see Appendix O) because the hypothesis test that $\hat{b}_1$ is equal to zero is the same as the hypothesis test that τ is equal to zero. The Kendall-Theil line has the desirable property of a nonparametric estimator: it is almost as efficient as the parametric estimator when all assumptions of normality are met, and is much better when those assumptions are not met (Helsel and Hirsch, 2003). A confidence limit for the slope of the line can be calculated by ordering the slopes m_{ij} for all $i < j;\ i = 1,2,\ldots,(n-1)$ and $j = 2,3,\ldots,n$ from smallest to largest, and selecting the r^{th} and s^{th} slopes such that the following inequality holds true:

$$P\left(m_{(r)} < M < m_{(s)}\right) \geq 1 - \alpha.$$

P-4.16. For more details about this confidence limit, see *Statistical Methods in Water Resources* (Helsel and Hirsch, 2003) or *Practical Nonparametric Statistics* (Conover, 1980).

P-5. Group Comparisons. In a manner similar to the comparison between background and on-site data, screening and definitive confirmation data can be compared as groups. After verifying that the minimum assumptions of the various tests are met, group means and variances can be

compared using *t*- and *F*-tests or their non-parametric equivalents (see Appendices M and N). In this case, the project team must decide on the decision confidence required, most likely α will be 0.2 or less. Methods for determining decision confidence levels are discussed in Appendix K.

P-5.1. The following provides a review of issues that must be considered when applying the method of group comparisons; the review primarily focuses on comparing distinctly different groups of data. Consider a site that contains areas of both high and low contamination. Given the extreme divergence in contamination levels, there will be different population means across the sampled areas. Sample data analyzed using Field Method A cannot simply be compared to the entire set of sample data using Laboratory Method B with a two-sample *t*-test (refer to Appendix N) because of the different mean levels of the measured contaminant. For this approach to be viable (i.e., two sample *t*-test based on field and laboratory methods), the underlying population would need to be relatively homogeneous. If this condition is not met, statistical tests for *paired data* would need to be used.

P-5.2. Paired statistical tests are recommended to determine whether Field Method A and Laboratory Method B are significantly different. To conduct these tests, an aliquot is homogenized and split into duplicates (it is possible the sample extracts would be split as well). One duplicate is analyzed by Method A and the other analyzed by Method B. For each data pair, the researcher evaluates the difference in results provided between Methods A and B. If the results from Method A are not different from corresponding results provided by Method B and the differences are normally distributed, then on the average, the difference between the two methods is zero. However, it should be noted that, as the differences are usually calculated over a *range of concentrations* (rather than at a single concentration), an average difference of zero does not necessarily demonstrate that Methods A and B are comparable. For example, it would be possible for Method A to produce much smaller values than Method B at low concentrations but much larger values at high concentrations so that, on the average, the differences between Method A and Method B over the entire concentration range is nearly zero. If Methods A and B are different, then the researcher should establish a functional relationship ($X_B = f(X_A)$) using regression analysis to "convert" the Field Method A results (X_A) to the corresponding laboratory Method B results (X_B) (see Paragraph P-4 for a discussion of regression analysis). The computed relationship, though, would need to quantify the uncertainty associated with the conversion. If this uncertainty is small relative to the uncertainty contributed by the field component, then the conversion uncertainty can be ignored and the "converted results" (X_B) used directly (i.e., can be treated as if they were directly obtained from a definitive laboratory method).

P-6. Percent Decision Match (PDM). The PDM may be a practical and useful approach to confirmation testing. The PDM is a qualitative evaluation strategy, as opposed to a more traditional statistical or quantitative strategy. For example, in the PDM, the decision error is not quantified and the variability in PDM results for a study area is not incorporated into the analysis. The PDM approach may be useful certain data quality objectives, namely to determine whether site contamination exceeds a specified decision limit.

P-6.1. The PDM is calculated as the number of times both data points in a data pair lead to the same conclusion divided by the total number of data pairs, expressed as a percentage:

$$\text{PDM} = \frac{Number\ of\ Decision\ Matches}{Number\ of\ Data\ Pairs}.$$

P-6.1.2. For example, suppose the regulatory threshold to which the data will be compared is fixed at 100 ppm. Suppose further that 100% of the data points from the screening technology are less than the threshold and the mean concentration is 50 ppm. Now, let us suppose that the definitive method of analysis systematically produces lower results and the mean concentration is 10 ppm. If both the screening data and the definitive data lead to the same conclusion, namely, that all of the samples are less than the threshold, is the difference between the absolute values of the screening and definitive analyses of any real significance? A PDM greater than 90% has historically been found to be acceptable to regulators in a number of differing jurisdictions.