STATISTICAL ANALYSIS PLAN
NORTHEASTERN POWER STATION
OOLOGAH, OKLAHOMA

Prepared in compliance with USEPA's Coal Combustion Residuals Rule, 40 CFR 257.93

January 2017

STATISTICAL ANALYSIS PLAN

Submitted to

1 Riverside Plaza
Columbus, Ohio 43215-2372

Submitted by

engineers | scientists | innovators

150 East Wilson Bridge Road
Suite 232
Worthington, Ohio 43085

In collaboration with

Kristina Rayner
Sanitas Technologies, Inc.

and

Kirk M. Cameron, Ph.D.
MacStat Consulting, Ltd.

CHA8423

January 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

Annual Report	Annual Groundwater Monitoring and Corrective Action Report
ANOVA	analysis of variance
CCR	coal combustion residuals
CFR	Code of Federal Regulations
GWPS	groundwater protection standard
LCL	lower confidence limit
MCL	maximum contaminant level
OLS	ordinary least-squares
ORP	oxidation-reduction potential
PQL	practical quantitation limit
QC	quality control
RCRA	Resource Conservation and Recovery Act
ROS	regression on order statistics
SAP	Statistical Analysis Plan
SSI	statistically significant increase
SSL	statistically significant level
SWFPR	site-wide false positive rate
Unified Guidance	*Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Unified Guidance* (USEPA, 2009)
UPL	upper prediction limit
USEPA	United States Environmental Protection Agency
UTL	upper tolerance limit

SECTION 1

INTRODUCTION

In April 2015, the United States Environmental Protection Agency (USEPA) issued new regulations regarding the disposal of coal combustion residuals (CCR) in certain landfills and impoundments under 40 CFR 257, Subpart D, referred to as the "CCR rules." Facilities regulated under the CCR rules are required to develop and sample a groundwater monitoring well network to evaluate if landfilled CCR materials are impacting downgradient groundwater quality. As part of the evaluation, the analytical data collected during the sampling events must undergo statistical analysis to identify statistically significant increases (SSIs) in analyte concentrations above background levels. A description of acceptable statistical programs is provided in USEPA's document *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Unified Guidance* (USEPA, 2009), which is commonly referred to as the "Unified Guidance".

The CCR rules are not prescriptive regarding what statistical analyses should be selected so that groundwater data are interpreted in a consistent manner and the results meet certification requirements. Geosyntec Consultants, Inc. (Geosyntec) prepared this Statistical Analysis Plan (SAP) on behalf of American Electric Power (AEP) to develop a logic process regarding the appropriate analysis of groundwater data collected in compliance with the CCR rules. The SAP will provide a narrative description of the statistical approach and methods used in accordance with the CCR rule reporting requirements [40 CFR 257.93(f)(6)].

This SAP describes statistical procedures to be used to establish background conditions, implement detection monitoring, and implement assessment monitoring (as needed) for the CCR units listed in **Table 1**. This SAP does not include statistical procedures for corrective action monitoring. These procedures should be developed when a corrective action groundwater monitoring program is established, if remedial action is necessary.

Procedures for collecting, preserving, and shipping groundwater samples are not included in this SAP. It is assumed that samples are collected and handled in accordance with AEP's draft *Groundwater Sampling and Analysis Plan* (AEP, 2016) and the requirements of 40 CFR 257.93 *et seq.*

SECTION 2

ANALYSES FOR REVIEWING AND PREPARING DATA

2.1 Physical Independence

Most statistical analyses require separate sampling events to be statistically independent. Statistical independence of groundwater samples is most likely to be realized when the samples are collected at time intervals that are sufficiently far apart that the samples are not from the same volume of groundwater. In such cases, the samples of groundwater are considered physically independent. To ensure physical independence, the minimum time between sampling events must be longer than the residence time of groundwater that would be collected in the monitoring well. The minimum time interval between sampling events (t_{min}) can be determined by calculating the groundwater velocity, as follows:

$$v = \frac{Ki}{n} \quad (1)$$

$$t_{min} = \frac{v}{D} \quad (2)$$

where:

$v =$ groundwater velocity
$K =$ hydraulic conductivity
$i =$ hydraulic gradient
$n =$ effective porosity
$t_{min} =$ minimum time interval between sampling events
$D =$ well bore volume (i.e., diameter of well and surrounding filter pack)

2.2 Testing for Normality

Many statistical analyses assume that the sample data are normally distributed. If such an analysis is used, the assumption of normality can be tested using the Shapiro-Wilk test (for sample sizes up to 50) or the Shapiro-Francía test (for sample sizes greater than 50). Normality can also be tested by less computationally intensive means such as graphing data on a probability plot. If the data appear not to be normally distributed (e.g., they are skewed in some fashion), then data may be transformed mathematically such that the transformed data do follow a normal distribution (e.g., lognormal distributions, Box-Cox transformations). Alternatively, a non-parametric test (i.e., a test that does not assume a particular distribution of the data) may be used. However, since non-parametric tests generally require large datasets to maintain an adequately low site-wide false positive rate (SWFPR), transforming the data is preferred.

2.3 Testing for Outliers

Outliers are extreme data points that may represent an anomaly or error. Data sets should be visually inspected for outliers using time series and/or box-and-whisker plots. While they are valuable as screening tools, visual methods are not foolproof. For example, if data are skewed according to a lognormal distribution, the boxplot screening may identify more outliers than actually exist. Typically, goodness-of-fit testing must be done on the non-outlier portion of the data to determine at what scale to test the possible outliers.

Potential outliers should be evaluated for potential sources of error (e.g., in transcription or calculation) or evidence that the data point is not representative (e.g., by examining quality control [QC] data, groundwater geochemistry, sampling procedures, etc.). Errors should be corrected prior to further statistical analysis, and data points that are flagged as non-representative should not be used in the statistical analysis. In addition, data points can be considered extreme outliers if they meet one of the following criteria:

$$x_i < \tilde{x}_{0.25} - 3 \times IQR \quad (3)$$

or

$$x_i > \tilde{x}_{0.75} + 3 \times IQR \quad (4)$$

where:

$x_i =$ individual data point
$\tilde{x}_{0.25} =$ first quartile
$\tilde{x}_{0.75} =$ third quartile
$IQR =$ the interquartile range $= \tilde{x}_{0.75} - \tilde{x}_{0.25}$

Extreme outliers may be excluded from the statistical analysis based on professional judgment. Goodness-of-fit testing may be needed to corroborate the classification of data points as extreme outliers. Flagged data and extreme outliers should still be maintained in the database and should be reevaluated as new data are collected.

2.4 Handling Duplicate or Replicate Data

Duplicate or replicate samples are often collected for QC purposes. Averaging the parent sample and duplicate sample results may give a more accurate representation of the constituent concentration at the time, but doing so would reduce the sample variability. Since many statistical tests assume that data are homoscedastic (i.e., the population variance does not change across samples), this technique is not recommended. Unless there is reason to suspect that either the parent sample or the duplicate sample is more representative of site groundwater, one of the samples should be selected at random and that value should be used in the subsequent statistical analysis. However, it should be reported when parent sample and duplicate sample results are

different from a decision-making perspective, e.g., when the duplicate sample exceeds the groundwater protection standard (GWPS) but the parent sample does not.

2.5 Handling Non-Detect Data

If non-detect data are infrequent (less than 15%), half of the reporting limit (RL) can be used in place of these data without significantly altering the results of a statistical test. The RL may be either the laboratory practical quantification limit (PQL) or an established project limit which is less than the maximum contaminant level (MCL). If non-detect data are more frequent, parametric methods that explicitly consider non-detects or non-parametric methods insensitive to the presence of non-detect data should be used. Where available, estimated results less than the RL (i.e., "J-flagged" data) should be used, and these data should be considered detections for the purposes of statistical analysis.

SECTION 3

DETECTION MONITORING

3.1 Establishing Background

By October 17, 2017, eight independent background samples should be collected from each monitoring well in the CCR unit groundwater monitoring system as part of the initial monitoring period [40 CFR 257.94(b)]. Background wells do not necessarily need to be hydraulically upgradient of the CCR unit, but they must not be affected by a release from the CCR unit [40 CFR 257.91(a)(1)]. The sampling frequency should be such that samples are physically independent, as described in **Section 2.1**. Samples should be analyzed for the Appendix III and Appendix IV constituents listed in **Table 2**.

Once analytical data are received, summary statistics (e.g., mean and variance) should be calculated for the background datasets. Initially, analysis should be done independently for each constituent at each well. As part of our protocol in such situations, time series plots and box plots will be prepared along with the summary statistics. The Kaplan-Meier method or robust regression on order statistics (ROS) can be used to compute summary statistics when there are large fractions (i.e., 15% to 50%) of non-detects; these methods are discussed below. If more than 50% of the data are non-detect, then summary statistics cannot be reliably calculated. Procedures for evaluating future data against these background datasets are described in **Section 3.2.1** (for detection monitoring) and **Section 4.1.1** (for assessment monitoring).

Background data will be evaluated for statistically significant temporal trends using (a) ordinary least-squares (OLS) linear regression with a *t*-test ($\alpha = 0.01$) on the slope and/or (b) the non-parametric Theil-Sen slope estimator with Mann-Kendall trend test ($\alpha = 0.05$). Non-detect data are replaced with half the RL for these analyses. The OLS linear regression or Theil-Sen slope estimator will be used to estimate the rate of change (increasing, no change, or decreasing) over time for each constituent at each well. The *t*-test or Mann-Kendall statistic will be used to determine whether a trend is statistically significant. OLS linear regression should only be used when at most 15% of the data are non-detect, when regression residuals are normally distributed, and when the variance from the regression line does not change over time. The Theil-Sen/Mann-Kendall analysis requires at least five observations for meaningful results; at least eight observations are recommended. Note that a statistically significant increasing trend in background data (or a statistically significant decreasing trend in pH) could indicate an existing release from the CCR unit or another source, and further investigation may be needed to determine the source of this trend.

If the trend analysis does not indicate a statistically significant trend, the proposed background data will be tested for normality using one of the methods outlined in **Section 2.2**. When data follow a normal or transformed-normal distribution (e.g. lognormal or other Box-Cox transformation), parametric methods are applied. If fewer than 15% of the data are non-detect,

non-detect data may be replaced with half the RL and the mean and variance can be calculated normally. If 15% to 50% of the data are non-detect, two methods – the Kaplan-Meier method or robust regression on order statistics (ROS) – can be used to determine the sample mean and variance. Kaplan-Meier should not be used if all non-detect data have the same RL or if the maximum detected value is less than the highest RL of the non-detect data. When data do not follow a normal or transformed-normal distribution, or when more than 50% of the data are non-detect, nonparametric methods may be used.

Once the sample mean and variance are calculated for each constituent at each well (assuming no significant trends over time), the data from background wells should be compared for each constituent. The purpose of this exercise is to test for significant spatial variation and to decide between interwell and intrawell approaches. First, the equality of variance across background wells should be tested visually using box-and-whisker plots and/or analytically using Levene's test (α = 0.01). If the variances appear equal, then one-way, parametric analysis of variance (ANOVA) should be conducted across background wells (α = 0.05). If there are no statistically significant differences between the background wells, then interwell comparisons may be appropriate to evaluate SSIs.

If ANOVA indicates statistically significant differences among background wells, then spatial variability can be concluded. As with temporal trends, the existence of spatial variability could indicate an existing release from the CCR unit or another source, and further investigation may be needed to determine the source of this variability. If the spatial variability is not caused by a release from the CCR unit, then intrawell comparisons would be appropriate to evaluate SSIs.

3.2 Evaluating Statistically Significant Increases (SSIs)

After the initial eight rounds of background sampling, groundwater sampling and analysis should be conducted on a semiannual basis. The statistical evaluation of each groundwater monitoring event must be completed within 90 days of receiving the analytical results from the laboratory [40 CFR 257.93(h)(2)].

The CCR rules only require analysis of the Appendix III constituents; however, analyzing additional parameters should be considered. Turbidity, dissolved oxygen, and oxidation-reduction potential (ORP), should be measured in the field in addition to pH. Other geochemical parameters, such as alkalinity, magnesium, potassium, sodium, iron, and manganese, should also be analyzed in the laboratory periodically (e.g., once every one to four years). Both the field and laboratory geochemical parameters can help identify the cause of any apparent change in groundwater quality. Additionally, analyzing for the Appendix IV constituents periodically should be considered to ensure the background dataset for these constituents is complete and current should assessment monitoring be needed. Statistical analyses should still be limited to the Appendix III constituents to help meet the dual goals of a SWFPR less than 10% per year and an adequate statistical power.

The CCR rules specifically list four methods acceptable for statistical analysis: ANOVA, tolerance intervals, prediction intervals, and control charts [40 CFR 257.93(f)]. Of these, the Unified

Guidance recommends prediction limits combined with retesting for maintaining a low SWFPR while providing high statistical power. Control charts are also acceptable as long as parametric methods can be used (i.e., the data or transformed data are normally distributed and the frequency of non-detects is at most 50%), as there is no nonparametric counterpart to the control chart. ANOVA is not recommended as the CCR rules mandate a minimum Type I error (α) of 0.05, at which it would be difficult to maintain an annual SWFPR less than 10%.

Prediction intervals and control charts can be used for both interwell and intrawell comparisons. For interwell comparisons, the pooled data from background monitoring wells should be used for the background dataset; for intrawell comparisons, the background dataset should be a subset of historical data at each monitoring well. (See **Section 3.4** below for procedures for updating background datasets.) Interwell comparisons are preferable, but they should only be used when there are no trends and no statistically significant population differences among background wells; otherwise, a significant test result may only indicate natural spatial variability instead of an SSI.

For prediction intervals, the upper prediction limit (UPL) is calculated according to the following formula:

$$\text{UPL} = \bar{x} + ks \quad (5)$$

where:

$\bar{x}$ = mean concentration of the background dataset
s = standard deviation of the background dataset
k = multiplier based on the characteristics of the site and the statistical test

Values for k are chosen to maintain an SWFPR less than 10% and depend on the following: (1) number of wells, (2) number of constituents being evaluated, (3) size of the background dataset, (4) retesting regime, and (5) whether intrawell or interwell comparisons are being used. Values for k are listed in Tables 19-1, 19-2, 19-10, and 19-11 in Appendix D of the Unified Guidance. If the k value that precisely matches site conditions does not appear in these tables, it can be estimated using the provided values by linear interpolation.

A one-of-two or one-of-three testing regime should be employed; i.e., if at least one sample in a series of two or three (respectively) does not exceed the UPL (or control limit), then it can be concluded that an SSI has not occurred. In practice, if the initial result does not exceed the UPL (or control limit), then no resampling is needed. If the initial result does exceed the UPL (or control limit), then a resample should be collected prior to the next regularly scheduled sampling event at the monitoring well(s) and for the constituent(s) exceeding the UPL (or control limit). Additional geochemical parameters, such as alkalinity, magnesium, potassium, sodium, iron, and manganese, should also be analyzed during resampling to help identify the source of the apparent increase. Enough time should elapse between the initial sample and each resample so that the samples are physically independent (**Section 2.1**). If both the initial result and the subsequent resample(s) exceed the UPL (or control limit), then an SSI can be concluded.

Choosing between a one-of-two and a one-of-three testing regime should be done before conducting the statistical analysis, as the UPL calculation depends on the resampling regime selected. The choice should depend on site conditions and the size of the background dataset. First, if three physically independent samples cannot be collected in a six-month period, then a one-of-two testing regime should be used. A one-of-two testing regime may also be considered (a) if the background dataset has at least 16 data points or (b) if the CCR unit's monitoring well network has nine or fewer downgradient monitoring wells and a background dataset of at least 8 data points. Otherwise, a one-of-three testing regime should be employed to achieve an acceptably high statistical power and an acceptably low SWFPR.

If two physically independent samples cannot be collected in a six-month period, then a reduced monitoring frequency may be warranted. In this case, a demonstration must be made documenting the need for – and effectiveness of – a reduced monitoring frequency. This demonstration must be certified by a qualified professional engineer, and monitoring must still be done on at least an annual basis.

The above procedure can be used wherever a mean and variance can be calculated for background data, including datasets that are transformed-normal and datasets where the mean and variance are calculated using Kaplan-Meier or Robust ROS methods. (Note that if data are transformed-normal, prediction intervals or control limits should first be calculated for the transformed data and then be transformed back into concentration terms.) Methods for determining prediction intervals where more than half of the background data are non-detect or where statistically significant trends exist are described below.

Different analyses can and should be used for different constituents and different monitoring wells within a CCR unit depending on the background data. For instance, if background wells have similar chloride data but different pH data, then interwell comparisons may be considered for chloride analysis and intrawell comparisons may be considered for pH analysis. If boron data are stable above the RL at MW-1, mostly non-detect at MW-2, and increasing at MW-3, then it would be appropriate to use parametric prediction limits at MW-1, to use non-parametric prediction limits at MW-2, and to conduct trend tests with potentially further investigation at MW-3 for boron at this site.

3.2.1 Most Background Data Are Non-Detect

If at least half of the data are non-detect, non-parametric prediction intervals with retesting should be used. In this method, the UPL is set either at the highest or at the second-highest concentration observed in the background dataset. A sufficiently large background dataset is paramount for this procedure to achieve an acceptably low SWFPR. To this end, the Kruskal-Wallis test should be performed on all background monitoring wells where at least 50% of the data for the constituent are non-detect to evaluate spatial variability. If the Kruskal-Wallis test indicates that there is no significant spatial variability among background wells, then the data from the background wells should be pooled to form a larger background dataset and thus to run an interwell test.

The choice between a one-of-two and a one-of-three testing regime should be based on the same criteria used for parametric testing, as described in **Section 3.2**. Choosing between using the highest or second-highest observed concentration as the UPL should depend in part on the size of the background dataset and the number of monitoring wells around the CCR unit. Assuming a one-of-three testing regime is used, the highest observed concentration should be used when the background dataset has fewer than 32 data points and the monitoring network has twelve or fewer wells. If there are at least thirteen wells, the highest observed concentration should be used when the background dataset has fewer than 48 data points. The second-highest observed concentration may be used for larger datasets.

If a one-of-two testing regime must be used due to aquifer conditions, then the highest observed concentration should be used (a) when the background dataset has fewer than 64 data points if there are fifteen or fewer wells or (b) when the background dataset has fewer than 88 data points if there are at least sixteen wells. The second-highest observed concentration may be used for larger data sets.

3.2.2 All Background Data Are Non-Detect

If all of the background data are non-detect, then the Double Quantification Rule should be used. According to this rule, if a sample and verification resample both exceed the PQL, then an SSI can be concluded. This can be thought of as setting the UPL at the PQL with a one-of-two testing regime. The possibility of false positives from this rule does not count against the calculated SWFPR because the false positive risk is quite small when all previous background data have been non-detect.

3.2.3 A Significant Temporal Trend Exists

True temporal trends in background data (i.e., absent a release from the facility or another source) are considered unlikely. Thus, a truncated dataset that does not exhibit a statistically significant trend may be used. In these cases, UPLs would be calculated as described in the previous sections.

Alternatively, if there is a significant temporal trend in the background data that is not attributable to a release, prediction limits can be constructed around a trend line. A trend line can be constructed parametrically using ordinary least-squares (OLS) linear regression. OLS linear regression should only be used when at most 15% of the data are non-detect, when regression residuals are normally distributed, and when the variance from the regression line does not change over time. If OLS linear regression is used, the UPL can be calculated according to the following equation:

$$\text{UPL} = \widehat{x_0} + t_{1-\alpha,n-2} * s_e * \sqrt{1 + \frac{1}{n} + \frac{(t_0 - \bar{t})^2}{(n-1)s_t^2}} \quad (6)$$

where:

$\widehat{x_0}$ = regression-line estimate of the mean concentration at time t_0
$t_{1-\alpha,n-2}$ = one-tailed *t*-value at a confidence of 1 – α and n – 2 degrees of freedom
s_e = standard error of the regression line
n = number of samples in the background dataset
t_0 = date the groundwater sample being compared to the UPL was collected
$\bar{t}$ = mean of the sampling dates in the background dataset
s_t = standard deviation of the sampling dates in the background dataset

The choice between a one-of-two and a one-of-three testing regime should be based on the same criteria used when there is no significant trend, as described in **Section 3.2**. The choice of α depends on the retesting regime and the number of wells within the monitoring network. If a one-of-two testing regime is employed, an α = 0.02 is recommended if there are eighteen or fewer wells and an α = 0.01 is recommended if there are at least nineteen wells within the monitoring network. If a one-of-three testing regime is employed, an α = 0.05 should be used.

3.3 Responding to an Identified SSI

If the statistical evaluation indicates that an SSI is present, the data should be evaluated to assess whether the SSI is caused by a release from the CCR unit. If it can be shown that the SSI resulted from a release from another source, from an error in sampling or analysis, or from natural variability, then a demonstration of this must be made in writing and certified by a qualified professional engineer within 90 days of completing the statistical evaluation. (The statistical evaluation itself must be completed within 90 days of receiving the analytical data from the laboratory.) If this demonstration is not made within 90 days of completing the statistical evaluation, then the site must begin assessment monitoring.

3.4 Updating Background

As recommended in the Unified Guidance, background values should be updated every four to eight measurements (i.e., every two to four years if samples are collected semiannually), assuming no confirmed SSI is identified. (See **Section 4.4** for procedures for updating background if an SSI has been identified.) A Student's *t*-test or the nonparametric Wilcoxon rank-sum test (also known as the Mann-Whitney test) should be conducted to compare the set of new data points against the existing background dataset, as appropriate. An α = 0.05 is recommended given the relatively small size of the datasets, particularly if background is updated every four measurements and particularly if the nonparametric Wilcoxon rank-sum test is used. However, an α as low as 0.01 may be used if the existing background dataset is large (e.g., if background data are pooled) or if Student's *t*-test is used.

If the *t*-test or Wilcoxon rank-sum test does not indicate significant differences, the new data should be combined with the existing background data to calculate an updated UPL. Increasing the size of the background dataset will increase the power of subsequent statistical tests.

If the *t*-test or Wilcoxon rank-sum test indicates a statistically significant difference between the two populations, then the data should be reviewed to evaluate the cause of the difference. Note that if the new data are significantly higher (or significantly lower in the case of pH) than the previous background data, this result would constitute evidence of an SSI and the response requirements outlined in **Section 3.3** would apply. If the differences appear to be caused by a release, then the previous background dataset should continue to be used and the site should enter assessment monitoring. Absent evidence of a release, the new dataset should be considered more representative of present-day groundwater conditions and used for background.

Once background datasets are updated, spatial variability should be assessed among background wells to determine whether pooling data and using interwell comparisons are appropriate, as outlined in **Section 3.1**.

SECTION 4

ASSESSMENT MONITORING

A CCR unit must begin assessment monitoring if an SSI is identified and is not attributed to some cause besides a release from the CCR unit. Assessment monitoring must begin within 90 days of identifying the SSI. During this 90-day period, the monitoring well network must be sampled for all Appendix IV constituents [40 CFR 257.95(b)]. Within 90 days of obtaining the results from this sampling event, all of the CCR unit wells must be sampled for all Appendix III constituents and those Appendix IV constituents that were detected during the initial assessment monitoring event [40 CFR 257.95(d)(1)].

After these initial assessment monitoring events, the CCR unit wells must be sampled for all Appendix III constituents and previously detected Appendix IV constituents on a semiannual basis [40 CFR 257.95(d)(1)]. Additionally, the CCR unit wells must be sampled for all Appendix IV constituents on an annual basis [40 CFR 257.95(b)]. Geosyntec suggests scheduling this annual sampling event to coincide with the required semiannual sampling.

As with detection monitoring, if physically independent samples cannot be collected on a semiannual basis, then a reduced monitoring frequency may be warranted. A demonstration must be made documenting the need for – and effectiveness of – a reduced monitoring frequency. This demonstration must be certified by a qualified professional engineer, and monitoring must still be done on at least an annual basis.

GWPSs must be established for each detected Appendix IV constituent. The GWPS shall be the greater of the background concentration and the MCL established by the USEPA for that constituent. If no MCL exists for that constituent, then the background concentration shall be the GWPS for that constituent [40 CFR 257.95(h)]. An upper tolerance limit [UTL] with 95% confidence and 95% coverage is often used as the representative background concentration.

A single site-wide GWPS would be recommended for each constituent based on pooled background data, even if natural spatial variability exists. If background data are not pooled, background concentrations and consequently GWPSs would vary from well to well. One difficulty with this approach is that concentrations at one monitoring well may exceed the location-specific GWPS and still be below levels considered as natural background at other locations within the site. The pooled background is often more interpretable and less cumbersome for developing a single background-based GWPS per constituent.

Once in assessment monitoring, two approaches can potentially be used to determine whether a move to corrective action is warranted. The more straightforward approach is to compare a confidence interval constructed on recent data against the GWPS. The GWPS is either set at the MCL or computed from background data (e.g., by calculating the UTL). When the lower confidence limit (LCL) of this interval exceeds the GWPS, corrective action may be justified.

The second approach only applies to cases where a published MCL does not exist, necessitating the use of background to determine regulatory compliance. In this case, an upper prediction limit on background data can be computed and compared against a predetermined number of recent measurements. The Unified Guidance recommends constructing a prediction limit for either a mean or median, depending on the background distribution, and then comparing the appropriate mean or median of the recent data. Exceedance of the prediction limit again may justify corrective action.

To return from assessment monitoring to detection monitoring, the CCR rules specify that all Appendix III and IV parameters must be at or below background levels for two consecutive sampling events [40 CFR 257.95(e)]. This can be tested by constructing a prediction limit on background for the next two future measurements for each parameter. If measurements from both sampling events do not exceed the prediction limit for every Appendix III and IV parameter, a return to detection monitoring would be warranted.

4.1 Comparing Data to the GWPS

As stated in **Section 4**, the GWPS is set at the MCL or (if an MCL does not exist for that constituent or if background data are higher than the MCL) a value based on background data. The UTL calculated from the background dataset is often used.

Tolerance intervals are similar to prediction intervals. However, whereas prediction intervals represent a range where a future result is expected to lie, tolerance intervals represent a range where a proportion of the population is expected to lie. Tolerance intervals have both an associated coverage (i.e., the proportion of the population covered by the tolerance interval) and an associated confidence. A coverage of 95% ($\gamma = 0.95$) and a confidence of 95% ($\alpha = 0.05$) are typically used.

The UTL is calculated similarly to the UPL:

$$\text{UTL} = \bar{x} + \tau s \quad (7)$$

Similar to the UPL calculation, $\bar{x}$ is the mean concentration and s is the standard deviation of the background dataset. However, in this case the multiplier τ is different from that of the UPL calculation and is a function of the chosen coverage and confidence and the size of the background dataset. Values of τ are tabulated in Table 17-3 in Appendix D of the Unified Guidance. As with prediction limits, if the τ value that precisely matches site conditions does not appear in these tables, it can be estimated using the provided values by linear interpolation.

Once a GWPS is established, new data must be evaluated to determine whether they are statistically significantly higher than the GWPS. The statistical analyses listed in 40 CFR 257.93(f) are appropriate for comparing new data to a background dataset, but are not appropriate for comparing new data to a fixed standard. For these cases, the Unified Guidance recommends using confidence intervals around the mean or median.

Evaluations should be done for each detected Appendix IV constituent at each well. Data from different wells should not be pooled. When selecting which data to include in the recent dataset, time series plots of concentration data at each well should be created and visually inspected. Only data that exhibit the same behavior as recent data should be included. For instance, if the last eight arsenic results cluster around 9 µg/L and the previous eight results cluster around 4 µg/L, then only the eight most recent results should be used in the statistical analysis. Similarly, if chromium concentrations steadily increased over the last ten samples and were stable previously, then the statistical analysis should only use the ten most recent results and (since they are steadily increasing) should involve constructing a confidence interval around a trend line.

At the same time, datasets should also be sufficiently large to maintain statistical power. As many data points that exhibit the same behavior as recent data as possible should be included, including data collected prior to assessment monitoring (e.g., during the initial eight monitoring events). Ideally, datasets should have at least eight data points; in no case should a dataset have fewer than four data points.

If at least 50% of the recent dataset is non-detect, then a parametric confidence interval should not be used, and the procedure in **Section 4.1.1** should be followed.

New data will be evaluated for statistically significant temporal trends using (1) ordinary least-squares (OLS) linear regression with a *t*-test ($\alpha = 0.01$) on the slope and/or (2) the non-parametric Theil-Sen slope estimator with Mann-Kendall trend test ($\alpha = 0.05$). Non-detect data are replaced with half the RL for these analyses. The OLS linear regression or Theil-Sen slope estimator will be used to estimate the rate of change (increasing, no change, or decreasing) over time for each constituent at each well. The *t*-test or Mann-Kendall statistic will be used to determine whether a trend is statistically significant. OLS linear regression should only be used when at most 15% of the data are non-detect, when regression residuals are normally distributed, and when the variance from the regression line does not change over time. The Theil-Sen/Mann-Kendall analysis requires at least five observations for meaningful results; at least eight observations are recommended. If a significant temporal trend exists, then a confidence interval around the trend line should be constructed as outlined in **Section 4.1.2**.

If the trend analysis does not indicate a statistically significant trend, then the mean and variance should be calculated. If fewer than 15% of the data are non-detect, then the non-detect data can be replaced with half the RL and the mean and variance can be calculated normally. Tolerance intervals are sensitive to the choice of population distribution. Normality should be confirmed using the Shapiro-Wilk (or Shapiro-Francía) test and/or probability plots, as described in **Section 2.2**. If data appear not to be normally distributed, data should be transformed so that the transformed data are normally distributed.

Two methods – the Kaplan-Meier method or robust ROS – can be used to determine the sample mean and variance when 15% to 50% of the data are non-detect. Kaplan-Meier should not be used

if all non-detect data have the same RL or if the maximum detected value is less than the highest RL of the non-detect data.

When most of the data are detections, data are normally distributed, and there is no significant temporal trend, the LCL is calculated according to the following equation:

$$\text{LCL} = \bar{x} - t_{1-\alpha,n-1} * \frac{s}{\sqrt{n}} \quad (8)$$

where:

$\bar{x}$ = mean concentration of the recent dataset
$t_{1-\alpha,n-1}$ = one-tailed *t*-value at a confidence of 1 – α and at n – 1 degrees of freedom
s = standard deviation of the recent dataset
n = number of samples in the recent dataset

The *t* value must be chosen in such a way to balance the competing goals of a low false-positive rate and a high statistical power. The Unified Guidance recommends that the statistical test have at least 80% power ($1 - \beta = 0.8$) when the underlying mean concentration is twice the MCL. Values of the minimum α (from which *t* values can be determined) are tabulated for this criterion for various values of *n* in Table 22-2 in Appendix D of the Unified Guidance. The selected α should be the maximum of the value in Table 22-2 and 0.01.

If data are transformed normal, the LCL should first be calculated for the transformed data and then be transformed back into concentration terms. Correction factors are available but are not expected to be required.

If the LCL exceeds the GWPS, then a statistically significant exceedance can be concluded. If this occurs, the owner/operator is required to take several actions, including potentially moving the facility to corrective action, as described in **Section 4.3**.

4.1.1 Most Data Are Non-Detect

If background data are mostly non-detect, non-parametric prediction or tolerance intervals should be used. In these cases, the UPL or UTL is set at either the highest or second-highest concentration observed in the background dataset. The highest or second-highest observed concentration effectively becomes the GWPS when this value is greater than the MCL. Even though the UPL and UTL are determined the same way, because of the difference in definitions between prediction intervals and tolerance intervals, the estimated statistical power and Type I error are different. The performance of the statistical tests will be evaluated, and the selected method will be based on the better statistical performance.

If recent data are mostly non-detect, non-parametric confidence intervals can be constructed around the median by ranking the data from least to greatest and setting the LCL equal to one of

the lower values of data. The confidence can be calculated based on the rank of the data point used and the sample size; confidence values are tabulated in Table 21-11 in Appendix D of the Unified Guidance for sample sizes up to 20.

However, if most of the recent data are non-detect, then the data point selected for the LCL will also be non-detect. If the RL is less than the GWPS, then no statistically significant exceedance has occurred.

GWPSs should only be determined for detected Appendix IV constituents [40 CFR 257.95(d)(2)]. If all the data for a constituent are non-detect, no statistical evaluation need be performed.

4.1.2 A Significant Temporal Trend Exists

If recent data show a significant temporal trend, then an LCL below the trend line can be calculated according to the following equation:

$$\text{LCL} = \widehat{x_0} - \sqrt{2s_e^2 * F_{1-2\alpha,2,n-2} * \left(\frac{1}{n} + \frac{(t_0 - \bar{t})^2}{(n-1)s_t^2}\right)} \quad (9)$$

where:

$\widehat{x_0}$ = regression-line estimate of the mean concentration at time t_0
s_e = standard error of the regression line
$F_{1-2\alpha,2,n-2}$ = upper (1 - 2α)th percentage point from an F-distribution with 2 and $n - 2$ degrees of freedom
n = number of samples in the recent dataset
t_0 = date of the most recent groundwater sample
$\bar{t}$ = mean of the sampling dates in the recent dataset
s_t = standard deviation of the sampling dates in the recent dataset

Note that the LCL is a function of time; to assess current compliance, the date of the most recent sample should be used for t_0. If and only if the LCL is greater than the GWPS at this time, then a statistically significant exceedance can be concluded. This equation can also be used to assess when the LCL will exceed the MCL (assuming the current trend continues).

The same α that would have been selected if there were no significant trend (as described in **Section 4.1**) should be used here to determine the proper F value.

If the Theil-Sen method is used to determine the trend line, a computationally intensive technique known as bootstrapping can be used to determine the LCL. This procedure is described in Section 21.3.2 of the Unified Guidance.

4.1.3 No MCL Exists

If no MCL exists, the GWPS can be set at the UTL calculated from the background dataset and compared to the LCL of the new data, as described above. Alternatively, a UPL can be calculated from the background dataset and compared to the mean or median of the new data. The GWPS is effectively set at the UPL. However, because the mean or median of the new data is being compared to this standard (as opposed to the LCL of the new data), the UPL standard is not directly comparable to an MCL or a UTL. Hence, this alternative should only be applied for constituents that do not have established MCLs.

As with UPLs with retesting presented in **Section 3.2**, UPLs can be calculated parametrically or non-parametrically. In the parametric test, the mean of the recent data is compared to the background UPL. The UPL is calculated according to the following equation:

$$\text{UPL} = \bar{x} + t_{1-\alpha,n-1} * s * \sqrt{\frac{1}{p} + \frac{1}{n}} \quad (10)$$

where:

$\bar{x}$ = mean concentration of the background dataset
$t_{1-\alpha,n-1}$ = one-tailed *t*-value at a confidence of 1 – α and at n – 1 degrees of freedom
s = standard deviation of the background dataset
p = order of (i.e., the number of samples used to calculate) the mean of the recent dataset
n = number of samples in the background dataset

As with other parametric tests, this test should only be applied if the data (or transformed data) appear normally distributed and at most half of the data are non-detect. Because the calculated UPL depends on the order of the mean of the recent dataset, the order of the mean (i.e., the size of the recent dataset) should be determined ahead of time. The choice of α will be made on a case-by-case basis to balance the goals of a high statistical power and low false-positive rate and will depend on the size of the background dataset (n), the order of the mean of the new dataset (p), and the number of wells in the monitoring network. Typical values of α range between 0.01 and 0.05.

If data are not normally distributed or if most of the data are non-detect, a non-parametric test can be run. In the non-parametric test, the median of the recent data is compared to the background UPL. The background UPL is set at either the highest or second-highest measured value in the background dataset. A sufficiently large background dataset is paramount for this procedure to achieve acceptably high statistical power and an acceptably low false-positive rate. To this end, background data should be pooled unless significant spatial variability exists. As with the parametric test, the order of the median should be determined ahead of time. The choices of using the highest or second-highest background data point as the UPL and of the order of the median will be made on a case-by-case basis. Selections will be made to balance the goals of a high

statistical power and a low false-positive rate and will depend on the size of the background dataset and the number of wells in the monitoring network.

In the rare cases where a significant temporal trend exists in background data and is not attributable to a release from the facility or another source, prediction intervals can be constructed around a parametric trend line, similar to the method outlined in **Section 3.2.3**. Alternatively, a truncated background dataset may be selected that does not exhibit a statistically significant trend and the UPL may be calculated as described above.

4.2 Comparing Data to Background

Assessment monitoring data must be compared to the GWPS to assess whether corrective action is warranted at the facility. Additionally, assessment monitoring data should be compared to background data to assess whether the facility can move from assessment monitoring back to detection monitoring.

To return from assessment monitoring to detection monitoring, the CCR rules specify that all Appendix III and IV parameters must be at or below background levels for two consecutive sampling events [40 CFR 257.95(e)]. However, the analysis of all Appendix III and IV parameters is not required for every monitoring event. Therefore, all Appendix III and IV parameters should be collected during two consecutive sampling events on a periodic basis (e.g., every two to four years) and/or when statistical evaluation of assessment monitoring data suggests groundwater concentrations are at or below background levels.

Prediction intervals of individual values or tolerance intervals can be used to compare assessment monitoring data to background. In either case, a UPL or UTL is calculated from the background dataset. Recent data are then compared to the UPL or UTL. If the measured concentrations are less than the UPL/UTL for every constituent at every monitoring well for two consecutive events, then it can be concluded that current concentrations are at or below background levels and a return to detection monitoring is warranted.

As described previously, prediction intervals and tolerance intervals can be constructed parametrically or non-parametrically. For the parametric intervals, the UPL is calculated according to Equation 5 and the UTL is calculated according to Equation 7. (Note that in this case, there is no retesting with the prediction intervals so the chosen k values will be different.) Non-parametric UPLs and UTLs can be determined by setting the UPL/UTL to the highest or second-highest measured background value. If all background data are non-detect, then future non-detect data can be considered statistically similar to background. If a temporal trend in background data exists and is not attributable to a release, prediction intervals can be constructed around the trend line or background data can be truncated so that no significant temporal trend is evident.

In all cases, the choice between prediction intervals and tolerance intervals, the choice of k or τ (for parametric intervals), the choice of the highest or second-highest background data point (for non-parametric intervals), etc. should be made based on sound statistical judgment and site

characteristics (e.g., size of background datasets, number of monitoring wells, etc.). For these statistical tests, because an SSI over background has already been identified, a higher α than that used in detection monitoring is justifiable.

4.3 Required Responses to the Results of the Statistical Evaluation

If the statistical evaluation demonstrates that the concentrations of all Appendix III and Appendix IV constituents are at or below background levels for two consecutive sampling events, then the CCR unit may return to detection monitoring [40 CFR 257.95(e)]. A notification that the CCR unit is returning to detection monitoring must be placed in the facility's operating record.

If the statistical evaluation demonstrates that some Appendix III or Appendix IV constituents are at concentrations above background levels but there are no statistically significant exceedances of GWPSs, then the CCR unit must remain in assessment monitoring [40 CFR 257.95(f)].

If the statistical evaluation demonstrates that an Appendix IV constituent is present at a statistically significant level (SSL) above its GWPS (i.e., if the LCL exceeds the GWPS), then the owner/operator must:

- Include a notification in the facility's operating record that identifies the constituents exceeding GWPSs [40 CFR 257.95(g)];
- Characterize the nature and extent of the release, including installing monitoring wells needed to delineate the plume, installing a monitoring well at the downgradient property boundary, quantifying the nature and the amount of the release, and sampling all wells for Appendix III and detected Appendix IV constituents [40 CFR 257.95(g)(1)];
- If the plume has migrated off-site, notify property owners overlying the plume [40 CFR 257.95(g)(2)]; and
- Either begin corrective action monitoring or demonstrate that the SSL is not due to a release from the CCR unit within 90 days of completing the statistical evaluation [40 CFR 257.95(g)(3)]. This demonstration must be made in writing and certified by a qualified professional engineer. The CCR rules require the previous three actions to be taken even if it can be demonstrated that the SSL is not due to a release from the CCR unit.

This SAP does not include statistical procedures for corrective action monitoring. These procedures should be developed when a corrective action groundwater monitoring program is established, if remedial action is necessary.

Reporting requirements for assessment monitoring are summarized in **Section 5.2**.

4.4 Updating Background

Care should be taken when updating background during assessment monitoring since, by definition, an SSI over background has already occurred. Data that appear to be affected by a release from the CCR unit should not be included in updated background datasets. However, it may be possible to update some background datasets (e.g., constituents not associated with a release, wells upgradient of the CCR unit, etc.). Formal updating of Appendix III parameters may be considered when there are at least four new points.

Data should be reviewed every four to eight measurements (i.e., every two to four years if samples are collected semiannually) to assess the possibility of updating background datasets. Professional judgment should first be applied; any data that appear to be affected by a release should be excluded from the background update, even if there is no statistically significant difference between the new data and the existing background data.

For data that appear not to be affected by a release, a Student's *t*-test or Wilcoxon rank-sum test should be conducted to compare the set of new data points against the existing background dataset. If the *t*-test or Wilcoxon rank-sum test corroborates that there are no significant differences, the new data should be combined with the existing background data to create an updated and expanded background dataset. Increasing the size of the background dataset will increase the power of subsequent statistical tests.

If the *t*-test or Wilcoxon rank-sum test indicates a statistically significant difference between the two datasets, then it should be assumed that the difference results from a release and the existing background dataset should continue to be used. If and only if there is evidence to suggest that the difference is not related to a release from the CCR unit, then the new dataset should be used for background.

Once background datasets are updated, spatial variability should be assessed to determine whether pooling data and using interwell comparisons are appropriate, as outlined in **Section 3.1**.

SECTION 5

REPORTING REQUIREMENTS

The CCR rule specifies reporting requirements throughout the monitoring process. Throughout the process, the required documentation is required to be posted both to the site's operating record and to a public internet set for review. As required by 40 CFR 257.93(f)(6), the chosen statistical methods described within this SAP are certified by a qualified professional engineer as appropriate for groundwater evaluation (**Section 6**). Reporting requirements relative to corrective action monitoring events are not discussed in this SAP.

By January 31, 2018, all existing facilities must submit an initial Annual Groundwater Monitoring and Corrective Action Report (Annual Report) [40 CFR 257.90(e)]. The Annual Report should be prepared and posted to both the site operating record and the public internet site. A notification should be sent to the State Director (and/or appropriate tribal authority) once the Annual Report is available.

The Annual Report should document site status, summarize key actions taken, describe problems encountered and their resolutions, and project key actions to be taken for the following year. The Annual Report should also include:

- A figure showing the CCR unit and the monitoring well network [40 CFR 257.90(e)(1)];
- An identification of monitoring wells installed or abandoned during the preceding year and the rationale for doing so [40 CFR 257.90(e)(2)];
- A summary of groundwater samples collected, which wells were sampled, what dates the samples were collected, and whether the samples were collected for detection monitoring or assessment monitoring [40 CFR 257.90(e)(3)]; and
- A discussion of any transition between monitoring programs (i.e., detection monitoring vs. assessment monitoring vs. corrective action monitoring).

If appropriate, the Annual Report should detail a demonstration for an alternative groundwater sampling frequency. If no SSIs are identified during each sampling event, an updated Annual Report should be submitted yearly.

5.1 Detection Monitoring

If SSIs are identified, the facility should demonstrate within 90 days of the detection, where possible, that SSIs over background are not due to a release from the facility, along with a certification by a qualified professional engineer that the information is accurate. If the SSIs over background are attributed to a release from the facility, the facility should prepare and place on the

operating record within 90 days a notification stating that an assessment monitoring program has been established [40 CFR 257.94(e)(3)].

5.2 Assessment Monitoring

If an assessment monitoring program is in place, the Annual Report must also include [40 CFR 257.95(d)(3)]:

- Analytical results for Appendix III and detected Appendix IV constituents,
- Background concentrations for all Appendix III and Appendix IV constituents, and
- GWPSs established for detected Appendix IV constituents.

The semiannual analytical results for Appendix III and detected Appendix IV constituents must also be posted to the facility's operating record within 90 days of receipt [40 CFR 257.95(d)(1)].

If a constituent is detected at an SSL above its GWPS, a notification must be reported to the site's operating record. Additionally, the facility must notify any person who owns or resides on land that directly overlies any part of an off-site contaminant plume and record the notifications in the facility's operating record. Within 90 days, the facility must either initiate an assessment of corrective measures or demonstrate that the SSL is not due to a release from the CCR unit. The demonstration must be supported by a report certified by a qualified professional engineer.

If statistics are performed by mid-October 2017 for the first compliance event, one or more resamples would normally be collected and re-analyzed within 90 days. By the end of January 2018, the initial exceedance will be either confirmed or determined to be a false positive. If it is confirmed, then assessment monitoring must be initiated within 90 days, which would fall at the same time as the next regular semi-annual event. In that case, the semi-annual event (March/April timeframe) would be for both assessment and detection monitoring (if assessment monitoring was initiated).

If the facility determines it may return to detection monitoring, the facility should issue a notification to the operating record and public site within 30 days.

SECTION 6

CERTIFICATION BY QUALIFIED PROFESSIONAL ENGINEER

By means of this certification, I certify that I am a qualified professional engineer as defined in 40 CFR 257.53, that I have reviewed this SAP, and that the statistical methods described therein are appropriate and meet the requirements of 40 CFR 257.93.

DAVID A. MILLER

Printed Name of Qualified Professional Engineer

David A. Miller

Signature

26057	OKLAHOMA	02.06.17
Registration No.	Registration State	Date

SECTION 7

REFERENCES

American Electric Power. 2016. Draft Groundwater Sampling and Analysis Plan. April 1, 2016.

Criteria for Classification of Solid Waste Disposal Facilities and Practices. 40 CFR §257. (2016).

Electric Power Research Institute. 2015. Groundwater Monitoring Guidance for the Coal Combustion Residuals Rule. Palo Alto, CA. 3002006287.

Environmental Protection Agency. 2009. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Unified Guidance. EPA 530/R-09-007.

Table 1

Applicable Coal Combustion Residual Units

Plant	Northeastern Power Station
CCR Units	CCR Landfill
	Bottom Ash Pond

Table 2

Monitored Constituents Under the CCR Rules

Appendix III to 40 CFR 257 – Constituents for Detection Monitoring

Boron
Calcium
Chloride
Fluoride
pH
Sulfate
Total Dissolved Solids (TDS)

Appendix IV to 40 CFR 257 – Constituents for Assessment Monitoring

Antimony
Arsenic
Barium
Beryllium
Cadmium
Chromium
Cobalt
Fluoride
Lead
Lithium
Mercury
Molybdenum
Selenium
Thallium
Radium 226 and 228 combined