NAC-0063_R0

Final Report on Updated Background Screening Levels

Plant Site, 1&2 SOEP and STEP, and 3&4 EHP
Colstrip Steam Electric Power Station
Colstrip, Montana

22 January 2016

Prepared for

TALEN MONTANA, LLC
P.O. Box 38, Colstrip, MT 59323

Prepared by

NEPTUNE AND COMPANY, INC.
1435 Garrison St, Suite 110, Lakewood, CO 80215

CONTENTS

FIGURES

TABLES

GLOSSARY

AOC	Administrative Order on Consent
bgs	below ground surface
BSL	Background/Baseline Screening Level
CCRAWP	Cleanup Criteria and Risk Assessment Work Plan
COI	Constituent of Interest
EHP	Evaporation Holding Pond
Ford Canty	Ford Canty & Associates, Inc.
Hydrometrics	Hydrometrics, Inc.
MDEQ	Montana Department of Environmental Quality
ML	Machine-learning
Neptune	Neptune and Company, Inc.
PAM	Partitioning around medoids
PPLM	PPL Montana, LLC
RF	Random Forests
SES	Steam Electric Station
SOEP	Stage One Evaporation Pond
STEP	Stage Two Evaporation Pond
Talen	Talen Montana, LLC
UCL	Upper Confidence Limit
USEPA	United States Environmental Protection Agency
UTL	Upper Tolerance Limit
WECO	Western Energy Company

Executive Summary

This report presents calculated baseline screening levels (BSLs) for evaluating potential groundwater impacts from the Colstrip Steam Electric Power Station (SES) located in Colstrip, Montana (the Facility). BSLs cover the Plant Site area (Plant Site), the Units 1&2 Stage I and II Evaporation Pond (SOEP and STEP) areas, and the Units 3&4 Effluent Holding Pond areas. Arcadis (2007) previously calculated BSLs for the Plant Site and Units 1&2 SOEP and STEP areas. Exponent previously calculated draft BSLs for the Units 3&4 Effluent Holding Pond (Exponent, April 18, 2011). Since that time considerably more data have been collected by various parties within the vicinity of the SES. Subsequent to approval of the BSLs Work Plan (Neptune and Company, Inc. 2015), data have been added to the available database, statistical evaluations have been performed, and revised BSLs have been calculated. The following objectives were achieved during the course of the data evaluation and BSL development:

1) Confirmed and updated the unimpacted status of wells (relative to SES closed loop wastewater operations) and groundwater samples from those wells used in previous developments of groundwater BSLs.
2) Identified additional wells that provide background data that were not previously included and evaluated them for inclusion in the groundwater background database.
3) Determined that the list of analytes with BSLs can be expanded based on the updated groundwater data.
4) Determined if BSLs are appropriate for site-wide use
5) Grouped stratigraphic units as possible, and practical, for BSLs calculation
6) Compiled and evaluated surface water data for exploratory data analysis and subsequent BSL calculation.
7) Updated statistical methodologies used in previous BSL calculation.
8) Presented updated BSLs.

Note that the BSLs Work Plan (Neptune 2015) offered some options for exploratory data analysis and methods for BSL calculations depending on the statistic of interest. Some clarification was provided after the BSLs Work Plan was approved; that is, the preferred BSLs were defined in consultation with the MDEQ as the 95th upper confidence bound on the 90th percentile of the baseline data. This statistic is often termed an upper tolerance limit (UTL), which, in this case can be written as a 95/90 UTL.

It was also noted in the BSLs Work Plan that the data would ultimately determine the statistical methods used for BSL calculations. Some options were offered in the BSLs Work Plan, but with the expectation that the data would drive the approach. In particular, it was noted that recent regulatory guidance would be followed, but that these methods would be augmented as necessary depending on the specifics of the available data. Various methods were considered, including some not described in the BSLs Work Plan, and, in consultation with the MDEQ, methods were agreed upon that are presented in this report. The sheer magnitude of the data led to consideration of some innovative methods for identifying background data, and data challenges (e.g., many non-detects, few data points) also led to using statistical methods that are fairly robust to such challenges.

For the large groundwater dataset, a Random Forests clustering approach is utilized to determine a baseline dataset from which BSLs can be estimated. BSLs are calculated for five different stratigraphic layers (Alluvium, Spoils, Clinker, Coal-Related, and SubMcKay).

For the smaller surface water dataset, the background data selected are those from four locations upstream of Colstrip, where the locations are chosen by subject matter experts, and choices are made based on sample location conditions, the number of sampling events, and the restriction that locations must be upstream from the Colstrip facility.

A bootstrapping method coupled with a Gehan-based ranking system to account for multiple detection limits within the non-detect data was used to estimate the background screening levels. This approach requires no assumptions about the distribution of the data.

Updated background screening levels are reported in Tables 7 and 9 for groundwater and surface water, respectively.

1.0 Introduction

Hydrometrics, Inc. (Hydrometrics), on behalf of Talen Montana, LLC (Talen; Formerly PPL Montana, LLC (PPLM)), retained Neptune and Company, Inc. (Neptune) to produce updated Background Screening Levels (BSL) for the Colstrip Steam Electric Power Station (SES) located in Colstrip, Montana (the Facility). The updated BSLs cover the Plant Site area (Plant Site), the Units 1&2 Stage I Evaporation Pond (SOEP), Stage II Evaporation Pond (STEP) area (1&2 Area), and Units 3&4 Evaporation Holding Pond (3&4 EHP) area.

On August 3, 2012, PPLM and the Montana Department of Environmental Quality (MDEQ) entered into an Administrative Order on Consent (AOC) Regarding Impacts Related to Wastewater Facilities Comprising the Closed-Loop System at the Colstrip SES (MDEQ/PPLM Montana, 2012).

As defined by the AOC, cleanup criteria for the constituents of interest (COIs) will be determined as follows (emphasis added):

> "For each COI in ground or surface water, except for the evaluation for ecological receptors, the applicable standard contained in the most current version of Circular DEQ-7 Montana Numeric Water Quality Standards ("DEQ-7"), the USEPA maximum contaminant level, the risk-based screening level contained in the most current version of Montana Risk-Based Guidance for Petroleum Releases, *whichever* is more stringent; and, for COIs for which there is not a DEQ-7 standard, a maximum contaminant level, or a risk-based screening level contained in the Montana Risk-Based Guidance for Petroleum Releases, the tap water screening level contained in the most current version of USEPA Regional Screening Levels for Chemical Constituents at Superfund Sites, **except that no criterion may be more stringent than the background or unaffected reference areas concentrations**; and
>
> For each COI in ground or surface water that may impact an ecological receptor, an acceptable ecological risk determined using the most current versions of standard USEPA ecological risk assessment guidance if the criteria set pursuant to (1) above are not adequate to protect ecological receptors, **except that no criterion may be more stringent than the background or unaffected reference areas concentrations**".

BSLs for groundwater have previously been established (Exponent 2011, Arcadis 2007, Maxim 2004). In the current work, Neptune presents updated BSLs for groundwater and surface water for the Colstrip SES. This Updated BSL Report is a companion to the Cleanup Criteria and Risk Assessment Work Plan (CCRAWP) currently being updated by FordCanty & Associates, Inc. (FordCanty) and Neptune based on initial comments from the MDEQ. As such, the results of the BSL statistical analyses presented here will be used to support the human health and ecological risk assessments.

1.1 Facility Description

A description of the Colstrip SES, hereafter called the Facility, is provided because BSLs need to be responsive to potential sources of contamination and to the different geologic and hydrologic conditions at the site. In particular, different groundwater datasets are developed for different stratigraphic layers. The transport mechanisms and the flow direction of groundwater between these layers can affect the background concentrations between some of these layers. In addition to environmental conditions, it is also important to understand the effect of existing groundwater capture systems. These have a significant influence on the flow of groundwater on and near the site, and limit flow directly downgradient of the capture systems.

The Facility consists of four power-generating units (Figure 1). Units 1 and 2 are 333 megawatts each and began operation in the mid-1970s. Units 3 and 4 are 805 megawatts each and began operation in 1984 and 1986, respectively. Talen is the operator of the Facility, which is co-owned by Talen, PacifiCorp, Puget Sound Energy, Inc., Portland General Electric Company, Avista Corporation, and NorthWestern Corporation (Hydrometrics 2015).

The Facility generates electricity through the combustion of coal. Fly ash, a by-product of coal combustion, is removed by air scrubber systems to reduce emissions. Bottom ash collects at the bottom of the boiler. Fly ash, bottom ash, and Facility wastewaters contain constituents of the original coal. A closed-loop process water/scrubber system is used at the Facility to minimize impacts to water resources in the area (the Facility is zero discharge). Ash- and water-based liquid wastes from the generating plants are impounded in ponds designed and constructed to control seepage losses. The Plant Site pond system includes ponds that serve all four generating units in various capacities. Fly ash disposal is not currently conducted on the Plant Site, but rather in holding ponds to the northwest of the Plant Site at the 1&2 Area and to the east of the Plant Site at Units 3&4 Effluent Holding Pond. Fly ash deposited during previous operations remains in the closed Plant Site Units 1&2 Pond A.

The Facility is located near the city of Colstrip, which lies within Rosebud County in the south central area of the state of Montana. Colstrip was established in the early 1920's to provide coal for Northern Pacific Railways locomotives. Mining ceased in the area in the late 1950's as diesel fuel replace coal as a fuel source for the locomotives. Mining resumed in the early 1970's to provide coal for the Colstrip Steam Electric Station, and other facilities. Coal mining, ranching, urbanization, and electrical generation are the primary land uses in the immediate Colstrip area.

Coal mining in the Colstrip area is accomplished by strip mining. This involves removal of the strata that overlies the coal, referred to as overburden. The overburden is blasted with explosives to make removal of the rock possible with the use of mining equipment. The coal is then typically blasted prior to removal. Following removal of the coal, the overburden from the next cut, is removed and placed in the pit. This material is referred to as spoil.

1.1.1 Geology

Stratigraphy in the Colstrip area consists of, from the surface downward, the Fort Union Formation, Hell Creek/Lance Formation, Fox Hills Sandstone, and Bearpaw Shale. The Fort Union Formation is divided into three members; the upper Tongue River Member, the middle

Lebo Shale Member, and the lower Tullock Member. The Tongue River Member is at the surface in the Colstrip area. The deeper Lebo Shale, and then the Tullock Members are exposed to the north. At Colstrip, the total thickness of the Fort Union Formation is about 650 feet. Figure 1 is a cross section that illustrates the geology in the Units 1&2 Stage I & II Evaporation Ponds area.

The Fort Union Formation consists of alternating and intercalated deposits of shale, claystone, mudstone, siltstone, sandstone, carbonaceous shale and coal. The formation was deposited in a fluvial system of meandering, braided, and anastomosed streams near the basin center and by alluvial fans at the margins. The fluvial systems were typically oriented northeast-southwest. (Flores and Ethridge 1985 as cited in Hydrometrics 2015).

Numerous coal seams are present in the Tongue River Member of the Fort Union Formation, the result of peat deposits that accumulated in swampy areas and channels. The main coal seams of interest near Colstrip are the sub-bituminous Rosebud (~ 24 feet thick) and McKay seams (~ 8-10 feet thick). The Rosebud Coal, however, is the only seam mined in the Facility area due to quality of the McKay Seam which makes it undesirable for use in many coal-fired boilers. Both the Rosebud and McKay coals contain natural vertical fracturing (cleats) generally oriented perpendicular to the bedding plane. Bedrock beneath the McKay coal stratigraphy is referred to as sub-McKay.

The Rosebud Coal, and in some places the McKay Coal has undergone *in situ* burning in the Colstrip area. Burned areas can be identified by red cap rock on hills around the region. Burning of the coal baked the overlying strata. As a result of the burning, the coal volume was reduced leaving a void for the overlying rock to collapse into, or slowly settle into over time. The thermally altered rock is referred to as clinker or scoria. Collapse of the rock resulted in secondary porosity (fractures). Permeability varies but is typically very high and depends on the amount of fine-grained sediments that has moved vertically into the available pore spaces and the degree and nature of fracturing. No clinker has been confirmed on the Plant Site proper but it does occur at the SOEP/STEP (Figures 2 and 3) and Units 3&4 Effluent Holding Pond areas.

Alluvium is present in the drainage bottoms. Figure 2 includes two cross sections illustrating the shallow geology along the Creek. The ancestral East Fork Armells Creek eroded through the shallow bedrock, including the Rosebud and McKay Coals, and in some places into the sub-McKay deposits (Figure 3).

1.1.2 Hydrology and Hydrogeology

Groundwater is found in multiple layers of stratum in the area. These include, in a general descending order:

- Fill – Typically earthen material that is used to fill depressions, backfill excavations or build up areas to create mounds or change the grade or elevation of the ground. Examples of fill are spoil placed back into a mine pit or standing on the edge of a pit, soil or aggregate placed in excavated areas, and fill placed to level roadways or parking lots, etc. Any disturbed soil that has been reworked, placed in another location, or disturbed and contoured would also be considered fill. In most cases, fill is above the groundwater table. However, in some instances, such as spoil, groundwater is present in the fill.

  - Spoil – Silt, clay, sandstone, coal fragments, formerly overburden units that have been used to backfill areas where the Rosebud Coal was mined. The spoil were formed as a result of strip mining of the Rosebud Coal seam. Strip mining involves removing overburden material (sedimentary rocks that overlie the coal) and placing it in the previously mined pit. The coal is then removed. The removed overburden is referred to as spoil. Groundwater flow directions in spoil are typically consistent with the area topography, or the orientation of the bottom of the pit until regional flow is re-established.

- Alluvium – Poorly sorted clay, silt, sand and gravel deposited by fluvial processes in drainage bottoms. The most significant alluvial deposits occur under East Fork Armells Creek, Cow Creek, South Fork Cow Creek, Stocker Creek, and Pony Creek. Groundwater flows down the drainages under gradients that are typically similar to the topography. Minor alluvial deposits are also present in tributaries. A basal gravel, comprised of clinker, is often present in the alluvium. Clinker fragments are typically also found throughout finer-grained alluvial deposits. Alluvium is usually saturated within a few feet of ground surface in the East Fork Armells Creek vicinity but may be unsaturated for all or part of the year in its tributaries and the upper reaches of the Cow Creek basins.

- Colluvium – Colluvium is slope deposits, which have been transported downslope by fluvial or gravitational means. Colluvium in the Colstrip area is most often a silty clay or clayey silt composition, although coarser deposits may be present locally. Colluvium is frequently inter-fingered with the alluvial deposits along the edge of floodplains. Groundwater is typically not present or is only present in small amounts in the colluvium.

- Rosebud Overburden – Bedrock units of the Fort Union Formation comprised of siltstone, claystone, shales, and fine-grained sandstone typically overlay Rosebud Coal. Groundwater is often present in the overburden units in the Plant Site Area and south of the Stage I Evaporation Pond area. Flow typically is in a direction similar to topography where groundwater is present.

- Rosebud Coal – Cleated coal with thickness on the order of 20 to 25 feet. This coal seam has been mined throughout much of the eastern portion of the Plant Site, south and southwest of the Stage I & II Evaporation Ponds, and west of the Units 3&4 Effluent Holding Pond.

  Groundwater levels (if present) in the Rosebud Coal drop as mining approaches, or pre-mining dewatering is conducted. Recharge of spoil groundwater begins once the pit is backfilled. Recharge is either laterally from adjacent coal (if the coal is wet), drainage into the spoil from adjoining overburden (if water is present), from infiltration of precipitation, or a combination. Additional information regarding groundwater flow can be obtained through review of recent annual hydrologic monitoring reports (Hydrometrics, 2015a), and in site specific AOC reports (Hydrometrics, 2013a, 2013b, 2015b). A detailed explanation of Colstrip, Montana Coal mining can be found in (Roberts et al, 1999).Groundwater flow in the coal is described in numerous permit documents for the Big Sky and Rosebud Mines.

  - Clinker – Also referred to as scoria and baked shale – Comprised of thermally altered and collapsed overburden (sandstone, siltstone, shale, etc.) formed by the burning of previously underlying coal. Clinker is generally quite permeable, a function of the secondary porosity caused by fracturing. Natural groundwater is typically not present in any of the three areas due to the high permeability that results.

- Rosebud-McKay Interburden. Typically consisting of siltstone and shale although isolated sandstone deposits may also be present. The thickness of the interburden, and the presence of groundwater varies throughout the area. The thickness typically ranges from less than one foot to more than 10 feet. Groundwater in the interburden generally flows in a direction similar to the Rosebud Coal.
- McKay Coal– Cleated coal with a thickness of 7 to 14 feet, but most often 8 to 9 feet. The McKay Coal is a widespread hydrostratigraphic unit in the Colstrip area as it is often saturated with groundwater. The McKay is absent, however, in areas along the western margin of the Plant site where it has been eroded, under much of the Stage I & II Evaporation Ponds, and in lower elevations in the Units 3&4 Effluent Holding Pond.
- Sub-McKay – Fort Union Strata consisting of interbedded claystone, siltstone, fine-sandstones, and thin coal seams. Channel sands are not uncommon. Multiple intervals of water bearing sandstone and siltstone are present. The shallower sub-McKay sandstone (first water under McKay Coal) is typically targeted for water supply wells. However, deeper intervals are also targeted in some areas where the shallower sands are dry or only contain limited amounts of groundwater or the shallower units have been removed by erosion. Channel sands are not uncommon. Sub-McKay sandstones are used for water supply aquifers in the Colstrip area. Yields from wells completed in sub-McKay sandstones in the Colstrip area vary from less than one gpm to more than 20 gpm.

Shallow groundwater flow directions are locally changed by the operation of current capture systems. For example, under non-pumping conditions at both the Plant Site and the 1&2 Area, shallow groundwater flow is generally expected to mirror the topography with flow towards the Creek and discharge into the alluvium along the Creek where the shallow bedrock units have been eroded by the ancestral East Fork Armells Creek. Under pumping conditions, overall shallow groundwater flow is locally diverted and interrupted by the capture systems (Figures 4 and 5). Groundwater flow is affected in a similar manner in the SOEP, STEP, and Units 3&4 Effluent Holding Pond areas.

Deep groundwater in the sub-McKay units generally flows to the northeast under a regional gradient toward the Yellowstone River.

Lateral variations in groundwater flow conditions might exist near mine spoil. If the hydraulic conductivity of the spoil is higher than the adjacent deposits, the spoil will act as a drain. Conversely, if the spoil hydraulic conductivity is lower, an impediment to flow will occur. Spoils are present in the eastern portion of the Plant Site. In general, permeability of the spoil is similar to the adjacent bedrock. However, spoil with a higher permeability is present north and west of the Units 3&4 Bottom Ash Ponds. This results in the high yield (~50 gallons per minute [gpm]) of the Western Energy Company (WECO) well. The WECO well was installed to lower the groundwater level below a coal crusher at the Rosebud Mine. The well was advanced to the base of the mine spoil (60 feet below ground surface [bgs]) and five feet into the underlying interburden (bedrock) to a depth of 65 feet. Spoil occurs west and southwest of the Units 3&4 Effluent Holding Pond but does not affect groundwater flow in the vicinity of the pond. Some active, or open, coal mine pits are also present. These pits act as groundwater drains when they intersect the water table.

Several indicator parameters are used to evaluate potential process wastewater impacts to groundwater at the Facility. These include specific conductance (SC), sulfate, dissolved boron, chloride, and the ratio of calcium to magnesium.

Existing groundwater capture systems in the areas where the highest concentrations of indicator parameters have been observed (both in the shallow units and in the McKay Coal) limit migration of impacted groundwater away from the Facility. At the Plant Site, capture wells are located downgradient of the Units 1&2 A Pond, Units 1&2 B Pond, Units 1&2 Bottom Ash Ponds, Units 1-4 Sediment Retention Pond, North Cooling Tower Blowdown Pond C, and South Cooling Tower Blowdown Pond C. Additional capture wells are located at the former Brine Ponds, Unit 3&4 Drain Collection Pond, and Units 3&4 Bottom Ash Ponds. Consequently, the Plant Site capture wells are located between the various ponds and East Fork Armells Creek. There is a small area with groundwater flow from the Plant Site toward the Cow Creek drainage basin near the Units 3&4 Bottom Ash Ponds. In the 1&2 Area, capture wells are located downgradient of the STEP dam between the dam and East Fork Armells Creek. In both locations the capture wells are designed to capture shallow groundwater prior to it reaching the creek. Groundwater capture is being conducted in the Units 3&4 Effluent Holding Pond in the alluvium downgradient from the Main Dam, the Saddle Dam, and in South Fork Cow Creek. Groundwater recovery is being conducted in the clinker along the south and southwest sides of the Units 3&4 Effluent Holding Pond. Groundwater recovery is being conducted from the sub-McKay sandstone directly north of the Units 3&4 Effluent Holding Pond and northeast of the pond.

1.1.3 Surface Water

East Fork Armells Creek

At the Plant Site and the 1&2 Area, the nearest natural surface water is East Fork Armells Creek. At the Units 3&4 EHP, the nearest surface water is Cow Creek. Regionally, the Creek is an intermittent stream, but it generally flows continuously through the town of Colstrip along the western edge of the Plant Site and along the eastern edge of the 1&2 Area. Surface water flow upstream and downstream of Colstrip is observed only in response to storm water or precipitation runoff events. Flow in the Creek varies throughout the year in response to runoff from precipitation, lawn watering, snowmelt, and plant growth. The Creek adjacent to the Plant Site and through the town of Colstrip is generally shallow and slow moving with abundant emergent aquatic vegetation present during the summer months.

At the Plant Site, the topography slopes downward from the Plant Site to the west/northwest toward the Creek. Colstrip SES is a zero-discharge facility, so there are no direct wastewater discharge points from the Plant to the Creek. Shallow groundwater from most of the Plant Site and the 1&2 Area flows in the direction of the Creek, though as discussed previously, a series of capture wells limit migration of groundwater to the Creek.

Water quality in the creek is affected by numerous activities and natural variations. These include but are not limited to:

- Influence from Castle Rock Lake,
- Influence from changes in runoff patterns to the creek due to industrialization or urbanization,
- Influences from development of sports facilities including ball fields and golf courses,
- Influence of runoff from the townsite that involves lawn maintenance, road maintenance, highway management, etc.,
- Influences from plant site capture systems and past seepage,
- Seepage from the City of Colstrip Treated Sewage Lagoons and storage ponds,
- Influence from upstream mining and interruption in surface water and groundwater flow to the creek.

Cow Creek, South Fork Cow Creek, Pony Creek
Other major drainages at the facility include Cow Creek and South Fork Cow Creek. These drainages are ephemeral in the headwaters. That is, there is only flow during response to snowmelt or precipitation runoff. Pony Creek is north of the Units 3&4 Effluent Holding Pond, and is also ephemeral. Water quality data are available from these drainages. However, the data are highly variable, and as such, are not considered useful for calculation of BSLs.

1.2 Previous Investigations

There have been two previous investigations of groundwater background conditions at the Plant Site and the 1&2 Area and one previous investigation at the Units 3&4 Effluent Holding Pond:

- A preliminary investigation of the 1&2 Area groundwater (Maxim 2004)
- A 2007 update of the Plant Site and 1&2 Area groundwater investigation (Arcadis 2007)
- A preliminary investigation of Units 3&4 EHP groundwater (Exponent 2011)

A preliminary statistical analysis of Plant Site and 1&2 Area groundwater data was conducted by Maxim Technologies in 2004 (Maxim 2004). Maxim identified a total of 59 wells in the area of the Plant Site and 1&2 Area that were deemed “unimpacted” by Facility operations and were included in the background analysis. The Maxim analyses divided wells into two groupings: “shallow” and “all wells.”

The statistical analysis previously performed by Maxim (2004) included the following steps:

1. Graphical analysis of the data distribution based on histograms, probability plots, and trend plots (scatter plots of concentrations against time).
2. Calculation of summary statistics, including the standard deviation, mean, median, minimum, maximum, range, and the sample sizes (including detects and non-detects). Non-detect (ND) values were taken to be the reporting limits (no substitution method was used such as half the reporting limit).
3. Calculation of BSL values based on the 95 percent upper confidence limit on the mean

(95 UCL) using a parametric method that assumes that data are normally distributed. The rationale for using the 95 UCL was attributed to USEPA's 1992 guidance, Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities (Addendum to Interim Final Guidance, July).
4. Trend plot analysis based on linear regression performed on concentration/time profiles.

Subsequent site investigations identified several areas for improvement in the identification of background groundwater samples for the Plant Site and SOEP/STEP areas. As a result, an effort was undertaken by Arcadis in 2007 to re-evaluate unimpacted wells identified by Maxim. This re-evaluation resulted in the removal of 18 wells and the addition of 33 others, bringing the total number of unimpacted wells to 74. Arcadis divided wells into three stratigraphic units (Bedrock, Alluvium, and Spoils). Wells that were completed in both the spoils and the bedrock were added to the Spoils dataset, and wells that were completed in both the alluvium and bedrock were added to the Alluvium dataset. The final unimpacted dataset evaluated by Arcadis included 15 Alluvium wells, 43 Bedrock wells, and 16 Spoils wells.

The Arcadis analyses evaluated 41 different analytes. Arcadis used the 95 percent confidence interval of the 95th percentile (95/95 upper tolerance limit, or UTL) to represent the BSL. However, sample sizes were considered sufficient for calculating 95/95 UTLs for only 16 analytes in Bedrock wells, and 4 analytes in Alluvial wells. No 95/95 UTLs could be calculated for Spoils wells. When sample sizes were not sufficient to calculate 95/95 UTLs, Arcadis used the maximum detected concentration (after outlier analysis) in the unimpacted wells to represent the BSL. Additional refinements to the Maxim analysis include the following:

1. Statistical procedures for identifying and testing outliers were added.
2. NDs were explicitly incorporated in the statistical analysis using non-parametric statistical approaches designed for left censored data. ND values were taken to be the reporting limits.
3. Trend analyses were conducted for the statistical evaluation of trends, including evaluations for seasonal cycles.
4. Additional analytes were included in the analyses to evaluate potential site impacts.
5. Suspect values were removed from the dataset prior to performing the statistical analysis (e.g., duplicate entries).

Additionally, an investigation by Exponent (2011) analyzed background groundwater conditions at Units 3&4 EHP by looking at samples taken prior to October 1, 1983. Results were presented in an external memorandum for the three stratigraphic units used by Arcadis (2007) with the separation of coal layers into a fourth unit called Coal. 95/95 UTLs were also used to estimate BSL values in this investigation, resulting in values for 23 analytes each in Alluvium, Bedrock, and Coal wells; and 37 analytes in Spoils wells. The statistical approach was similar to Arcadis (2007) in that outlier tests were performed, non-parametric UTLs were calculated, and the maximum detected concentration was used as the BSL when sample sizes were insufficient for bootstrapping.

Groundwater well data have been collected for more than seven years since the time that BSLs were last calculated for the Plant Site and SOEP/STEP areas and five years since preliminary BSLs were estimated for the Units 3&4 EHP. In addition, the conceptual model of the site continues to be improved as more information becomes available. The current investigation includes an assessment of the wells that were used in the initial evaluation, as well as the updating of the BSLs developed in earlier investigations to include new data and potential data that was previously undiscovered.

1.3 Summary of Current Approach

Draft groundwater BSLs have been calculated previously for the Units 3&4 Effluent Holding Pond, but final BSLs have not been promulgated by MDEQ. In this study site-wide data were used, including data from the Units 3&4 Effluent Holding Pond. An additional component of the investigation was evaluation of stratigraphic layers used for BSL calculations.

The possibility of developing BSLs for surface water was also evaluated, since earlier investigations did not address surface water. East Fork Armells Creek has been sampled since the mid-1980s, but sampling locations include only four locations that can be considered upstream of the Plant's influence. This is because upstream of the town of Colstrip, East Fork Armells Creek experiences intermittent flow. Part of the current investigation was to explore additional upstream locations and calculate surface water BSLs if enough data existed.

The broad objectives of the current investigation are as follows:

- Confirm and update the unimpacted status of wells (relative to SES closed loop wastewater operations) and groundwater samples from those wells used in previous developments of groundwater BSLs
- Identify additional wells that provide background data that were not previously included and evaluate them for inclusion in the groundwater background database
- Determine if the list of analytes with BSLs can be expanded based on the updated groundwater data
- Determine if BSLs are appropriate for site-wide use
- Group stratigraphic units as possible, and practical, for BSLs calculation
- Compile and evaluate surface water data for exploratory data analysis and subsequent consideration for BSL calculation
- Update statistical methodologies used in previous BSL calculation
- Present updated BSLs

The development of updated groundwater and surface water BSLs started with the formation of a suitable dataset, involving additional samples gained since the previous investigations and additional sources of data, such as the Montana Bureau of Mines and Geology. For the groundwater dataset, a machine learning (ML) approach called Random Forests (RF) was adopted in order to use modern computational abilities to divide the dataset into two groups: samples not used in BSL calculations (non-baseline) and samples used in BSL calculations (baseline). ML is a broad area of computer science used to understand and model large, complex datasets. ML methods are better suited than traditional statistical methods for analyzing the Facility groundwater data because ML methods provide efficient means for handling large datasets (here, more than 600,000 groundwater samples) with many variables (over 200 different analytes appear as measured quantities in the samples) and with a high degree of omission (many samples have measurements for only three or four dozen analytes, for instance). All three of these data traits (large datasets, many variables, and many missing data values) create obstacles of various types for traditional methods. The ML method used is called Random Forests (RF). The random forest method is robust to extreme values and to skewness in the data distributions. These qualities are important because the concentration distributions for many analytes are strongly right-skewed.

The RF methodology is used here to determine sample similarities. These similarities are then used in a clustering algorithm to divide the data into the two groups. RF has been used in a wide variety of environmental applications, including predicting tree species distribution (Mellor *et al.* 2013, Prasad *et al.* 2006, Evans & Cushman 2009), forest carbon stores (Mascaro *et al.* 2014), air temperature (Ho *et al.* 2014), ecological classifications (Cutler *et al.* 2007), and groundwater quality (Rodriguez-Galiano *et al*. 2014, Anning *et al.* 2012). RF has also been used in predicting medical diagnoses (Wolfe *et al.* 2010) and gene selection (Díaz-Uriarte & Alvarez de Andrés 2006). Further details about the method are provided in Section 4 and Appendix A. Additionally, a simplified example of the RF clustering approach applied in this investigation is provided in Appendix B.

The focus of the current approach is on samples, not sample locations (i.e. wells). A single sample can be identified by its sampling date and sample location and usually has results for multiple analytes. Using the sample as the unit of evaluation allows for only a part of a well's historical sampling record to be used in BSL calculations. In other words, if the sampling record of a well shows effects of contamination in the latter half of the record, the first half of the record could still be used for BSL calculations. However, once a well showed evidence of impacts, no samples at later dates are used.

Each sample is treated independently and is not assumed to have a relationship with past or future samples taken from the same well. Therefore seasonal and/or temporal trends are not directly assessed as they would be in a more traditional statistical analysis.

The goal of the RF approach is to classify each sample as baseline or non-baseline in a manner representative of theoretical envelopes of expanding spatial and unidirectional temporal impact, and to rule out samples classified as elevated compared to baseline samples in the development of BSLs. Although RF is unlikely to identify the spatial and temporal boundary perfectly, this statistical approach proved to be a very useful first step at identifying background samples from the very large dataset that had been compiled. More traditional approaches to statistical clustering have difficulty with such large datasets, especially when the data are highly skewed. With the large amount of data available for the purpose of estimating BSLs, it is not necessary to identify every background sample; it is necessary only to identify sufficient background samples to support BSL calculations. RF efficiently combed through the data to separate baseline and no-baseline data that are then compared to the history and conceptual model of the site.

The baseline groundwater dataset resulting from the RF clustering approach consisted of samples defining groundwater conditions unimpacted by SES operations. This approach is used to identify two basic types of samples that represent 'baseline' conditions for the Facility. The first type included on-site and downgradient samples prior to Units 3&4 Effluent Holding Pond pond construction (ca. 1983) and samples prior to or outside the spatial range of impacted groundwater flow from process ponds. The second type are samples upgradient of the Facility that may have been affected by activity other than Facility operations but nonetheless have the potential to be present in future Facility samples depending on groundwater flow. The first type of samples represent on-site background conditions while the second type represents upgradient baseline conditions. Baseline conditions affect on-site groundwater quality but are not under the control of SES operations. The RF clustering approach results in a dataset that is a combination of both types of samples, but that is referred to here as a baseline dataset. BSLs were based on this baseline dataset. Samples not included in the baseline dataset are termed ‘non-baseline’ and represent samples that may have site impacts.

A brief overview of the current approach in regards to groundwater:

1. Data preprocessing to combine data from various sources.
2. Identification and separation of groundwater and surface water sampling locations.
3. Determination of sample similarities using the RF machine-learning algorithm.
4. Clustering of samples into baseline and non-baseline groups based on the RF-determined similarities.
5. Expert review of clusters and refinement of baseline and non-baseline assignments.
6. Calculation of BSLs based on unimpacted samples.

The approach for surface water was different, because surface water samples are limited to a relatively small number of locations on East Fork Armells Creek appropriate for inclusion in the baseline dataset. Specifically, locations were limited to those upstream of the Facility with enough samples and continuous flows. Because of the small number of potential sample locations, samples were assessed on a location-by-location basis, and samples from four locations were identified for use in baseline surface water calculations.

2.0 Data Preprocessing

Water quality data from both groundwater and surface water sites were provided by Hydrometrics, Inc. (Hydrometrics) as Excel (Microsoft) spreadsheets from several sources: Talen Montana LLC, Rosebud Mine (Westmoreland, formerly Western Energy), Big Sky Mine, Battelle, and the Montana Bureau of Mines and Geology (2015). These data were combined into a single dataset and stored as a data table in a PostgreSQL database (PostgreSQL 9.3 PostgreSQL Global Development Group). To combine the data from various sources, the spreadsheets were read into the open-source statistical software R (R Core Team 2015) as comma-delimited files and processed to standardize column names and field entries (such as units, analyte names, stratigraphic layer, well purpose, and detect flags). Additional columns were added to distinguish groundwater samples from surface water samples and to incorporate well metadata, such as status, found in other files. In total, this resulted in a dataset of 641,793 samples from 2,206 wells for 285 analytes over a timeframe spanning from August 8, 1972 to June 30, 2015.

2.1 Development of a Groundwater Dataset

To develop a groundwater dataset, surface water locations are removed. Next, to ensure data are representative of the environment in which the Facility is located, a spatial constraint is applied to omit data west of the Rosebud Mine, south of the Big Sky Mine, north of Pony Creek, and east of the confluence of Cow Creek with South Fork Cow Creek. Samples from process sites, dam sumps, interception trenches, test holes, boring holes, and pits are removed. Also removed are dry wells, wells with evidence of process water impacts, wells at poor locations (directly down gradient of dam collection sumps, for example), wells with no logs or questionable completion information, and/or wells with histories that made their water quality record suspect. One such example is well PSW-1, which was perforated from the bottom up during the original installation. Remaining wells were reviewed with Hydrometrics on a well-by-well basis, along with bore logs and/or construction information, to confirm their stratigraphic layer and suitability for inclusion in further analysis. A majority of wells reviewed at this point are monitoring or capture wells. Wells without construction information, without well logs, or with information that suggested they were completed over more than one layer are filtered out. Wells from eleven different stratigraphic layers remained: Alluvium, Colluvium, Shallow, Spoils, Clinker, Rosebud Overburden, Rosebud Coal, Interburden, McKay Coal, SubMcKay, and SubMcKay Deep.

For groundwater data, these steps resulted in a dataset with 145 unique analytes. However not all analytes had enough data to be carried further, and so this dataset is further reduced to 47 analytes based on data availability, including all analytes from the most recent previous investigation (Arcadis 2007) except field conductivity plus seven additional analytes. The RF and clustering methods are applied to these analytes. Analytes not carried forward are summarized by data availability in Appendix I. For groundwater data, these steps ultimately resulted in a dataset with 356,297 samples from 1,333 wells, with sampling dates from March 23, 1973 to June 30, 2015. Nearly all groundwater metals samples are filtered. Data availability by analyte and stratigraphic layer can be found in Table 1.

2.2 Development of a Surface Water Dataset

The first step in the creation of a surface water dataset for BSLs estimation is to rule out samples from sampling locations previously identified as groundwater locations. There were additional types of sampling locations also ruled out as a second step, such as springs and mine or city outfalls. This was further narrowed down to only sampling locations on East Fork Armells Creek, which is the surface water source that flows through the Site, and of those locations, only sampling locations upstream of the Plant Site were kept. Additional remaining locations were ruled out if they were ephemeral, run-off, out falls, springs, seeps, or ponded water and based on discussion with MDEQ personnel. The final surface water dataset contained four locations, 1,684 samples, 39 of the 47 analytes considered in the groundwater dataset (no antimony, nitrite, nitrate, phosphate, titanium, silica, silver or tin), and a temporal span of February 14, 1981 to October 16, 2014. A majority of surface water metals are unfiltered, and most essential nutrient samples are filtered. Data availability by analyte can be found in Table 2.

3.0 Exploratory Data Analysis

Because the goal of the analysis is the development of site-wide BSLs, the data are not split into site sub-area specific datasets. However, the groundwater data are split into separate datasets based on stratigraphic layers. Initially, eleven layers are considered (Alluvium, Colluvium, Shallow, Spoils, Clinker, Rosebud Overburden, Rosebud Coal, Interburden, McKay Coal, SubMcKay, and SubMcKay Deep). Examination of boxplots (Figures 6-11) for six indicator analytes (specific conductance (SC), sulfate, dissolved boron, chloride, and the ratio of calcium to magnesium) suggested that further groupings for these layers is possible:

- Alluvium, Colluvium, and Shallow are combined into one stratigraphic unit, called Alluvium. Wells previously excluded because they bridged the Alluvium/Colluvium layer are included in this subunit of data as well.
- SubMcKay and SubMcKay Deep are combined and called SubMcKay.

Only adjacent stratigraphic layers are considered for combination. For example, SubMcKay is not considered for combination with Alluvium. The decisions to combine layers are based on a visual inspection of boxplots (Appendix C). This results in eight stratigraphic units, each of which is subjected to the RF clustering process. Boxplots, histograms, and Q-Q plots[1] for each of the resulting eight combined stratigraphic units can be found in Appendix D. In all plots non-detects are plotted at their reported detection limits using a hollow circle while detected values are plotted using a filled circle. This differentiates non-detects from detects, while maintaining the actual reported values for visualization of the data. The number of dimensions to the dataset and the large number of data points for many analytes make traditional statistical significance comparisons ineffective (traditional statistical tests with large number of data points tend to identify too many statistically significant results). Consequently, trend tests and other types of statistical comparison tests have not been performed. In light of the goal of the analysis, to identify samples that represent background conditions, the random forests approach is more adept at handling large and multi-dimensional data.

Because the groundwater dataset spans over 40 years, detection limit values are also plotted over time in order to examine the potential for changes in analytic techniques over time to affect analyses (Appendix D). No obvious patterns in detection limits over time are apparent.

Surface water concentrations over time are plotted by analyte and sampling location for each of the four selected locations (AR-12, SW-55, SW-60, and SW-75; Appendix E). Time series plots, Q-Q plots, and histograms are presented by analyte (Appendix E). Conceptually, all samples from these locations represent upstream baseline conditions unimpacted by SES operations. Some trends in concentrations are suggested in the time plots, however, there is sufficient variability in the concentrations that the trends might be an indicator of seasonal and annual effects on stream water flow. That is, high flow might be associated with lower concentrations. Consequently, the following highlights should be considered with some caution.

- AR-12 and SW-55 exhibit some change in concentration over time.

[1] **Boxplots** are a method of representing the distribution of a dataset. The top and bottom of the box in the boxplot represent the Inter-Quartile Range (IQR), identified by the 75th, and 25th percentiles of the data, respectively. The horizontal line in the middle of the box represents the 50th percentile (the median). Vertical lines (called whiskers) extend to last data point which is no more than 1.5*IQR from the box. Data points beyond the whiskers are represented by circles. A **Q-Q plot** is a way to check how normal a dataset is. It involves graphing the quantiles of a data set against the quantiles of the standard normal probability distribution. If the data are normally distributed, then the plotted pairs will follow a straight line. **Histograms** plot the frequency of observations within consecutive, equally sized intervals of concentrations. They provide a discrete estimate of the shape of the distribution of a dataset.

  - concentrations of calcium, chloride, magnesium, manganese, potassium, sodium, and sulfate potentially decrease over time for AR-12
  - concentrations of these same analytes appear to increase over time for SW-55
- Most of the samples at SW-60 and SW-75 are collected over shorter time periods and exhibit considerable variability.

Overall these plots suggest that it is reasonable to use all of the samples available at all four locations for developing BSLs. A tabular summary of this baseline dataset is provided in Appendix H and a map of its spatial extent is provided in Figure 25.

4.0 Identification of Baseline Groundwater Data

For each of the eight stratigraphic groundwater datasets, three basic steps are involved in establishing the baseline groundwater dataset for that unit. The first two steps are statistical – first to apply the RF algorithm (Breiman 2001, Liaw & Wiener 2002), and second to use the RF output to cluster samples into two groups, differentiating background concentrations from elevated concentrations. The third step involves a sample-by-sample review of the clusters to ensure that BSLs represent baseline conditions (on-site and upgradient) rather than only background conditions.

RF is used to calculate a similarity measure for samples based on concentrations, or values, for their multivariate suite of analytes (see Appendix A for details). The RF analysis uses 39 analytes instead of the 47 mentioned above. Field-measured specific conductance is considered too highly correlated with lab-measured specific conductance for separate inclusion. RF is run on 39 analytes to identify background samples. Once those samples are identified, BSLs are calculated for all 47 analytes. Qualities of the RF process that make it especially appropriate for this purpose are its insensitivity to extremes in the dataset and its ability to compare variables with different ranges and units (scale invariance). In addition, RF does not require data to be normalized or to fit a particular statistical distribution, or to be transformed in any way.

Because some analytes are measured much less frequently than others, RF starts by filling in missing values using a process called RF imputation. This RF imputation process is slightly different from the RF process used to generate the clustering input (similarity values), but uses the same core RF algorithm. RF imputation uses information from other samples to fill in realistic values for missing data. A more detailed description of RF imputation is found in section 4.1 and Appendix A. The performance of this method is tested by leaving out non-missing data, imputing values for the omitted data, and then comparing the imputed and known values. Based on this evaluation, analytes with less than 500 non-missing values are not used in the initial clustering assignments of samples to impacted and unimpacted groups. Table 3 shows the analytes dropped from each stratigraphic unit for this step. Imputed values are used only to evaluate the similarity of samples for clustering purposes. They are not used to calculate BSLs.

After RF imputation, another RF process is used to generate a sample similarity matrix (technically, a “proximity” matrix; for details see Appendix A) for each stratigraphic unit. RF also produces estimates in the relative importance of variables, in terms of how much each variable contributes to the overall structure of the data. Relative importance plots are presented in Appendix F, although they do not directly impact clustering results. Factors other than analytic concentrations can be included in RF, and time and spatial coordinates were included as additional input variables in preliminary RF dissimilarity runs. However, time (date of sampling) and spatial coordinates were not important predictors of similarity and are subsequently left out of the final RF analysis. This is not necessarily advantageous or disadvantageous, but instead indicates that the concentration differences alone inherently capture the spatial and temporal distinctions.

The degree of similarity produced by the RF is captured as similarity values for each pair of samples. The similarity values are presented in a matrix. The similarity values are then used to find groups (clusters) of samples that include similar data using an approach called "partitioning around medoids" (PAM). The PAM clustering algorithm clusters samples that have high similarity values, while separating those with low similarity into different clusters. The PAM method clusters by finding a group of samples that are near a medoid, which is a point in the cluster whose average dissimilarity to all objects in the cluster is minimal. So, in effect, background samples will appear similar, and non-background samples will appear similar, but background samples will not appear similar to non-background samples.

PAM results for two, three, and four clusters are created and plotted for the six most important analytes as determined by RF variable importance scores (Appendix F). After visual inspection, the clustering results for only two medoids are used. Additional clusters provide no better delineation between clustering groups (i.e., background and non-background).

The PAM clustering process captured the bulk of the classification of background and non-background samples. However, experienced hydrogeologists familiar with the history and geology of the Facility and its surrounds reviewed the statistically-defined clusters to evaluate if some samples should be moved from non-background to background and vice versa. In part this step involved consideration of upgradient baseline conditions as opposed to background conditions. That is, some samples identified as non-background are upgradient of the Facility and can be classified as background because they are considered a baseline condition for the Facility. Classifying these statistically would be very difficult, but classifying them based on known site conditions, site history, groundwater flow, and knowledge of local and site hydrogeology was considered reasonable. This expert review was also used to evaluate wells and individual groundwater samples that could have been misclassified statistically. That is, for some samples the classification is clear, but for others the classification is more uncertain (and additional clusters did not help define this). The RF process averages thousands of iterations (Appendix A), and will not necessarily result in grouping borderline samples in the same clusters over subsequent runs due to the random nature of RF. In addition, there is always uncertainty in the data, which could lead to misclassification of some samples. The expert review adds another level of conformity to clustering results. Furthermore, not every background sample needs to be identified, only a sufficient number to support BSL calculations.

The result of the RF imputation, RF dissimilarity (similarity – see Appendix A), PAM clustering, and expert review processes is a collection of baseline samples (on-site background plus baseline upgradient) that are used for BSL determinations. The spatial distribution of wells that do or do not fall into this baseline dataset, along with wells that have samples both in an out of it, can be seen in Figure 18. A more detailed description of the steps taken to identify the baseline dataset is provided below (sections 4.1 through 4.4) and in Appendix A.

4.1 Random Forests Imputation

As noted above the first step in the RF cluster analysis is to fill in or "impute" missing data. The imputed values are not used in computing BSLs, but are necessary for maximizing the amount of data available for the RF dissimilarity analysis (see Section 4.2). Although imputed values add no new information (they are calculated entirely from non-missing data), they benefit the analysis by allowing all of the non-missing data to inform RF dissimilarity.

RF is a machine-learning algorithm that makes use of modern computing power to iteratively identify relationships among observations of multivariate data. It can be used to classify samples from unknown groups based on patterns in samples with known groups (i.e. supervised RF), and it can also be used to partition samples into groups when no true group assignments are known (i.e. unsupervised RF). RF imputation is a supervised type of RF used to replace missing values based on the relationship between known values, effectively imputing much like parametric regression methods, but often with better results than parametric or mixed-type methods (Shah *et al.* 2014, Stekhoven & Buehlmann 2012).

For each of the eight stratigraphic units of input data, a separate imputation was performed using an iterative RF imputation process in which missing values are estimated and re-estimated until the sequence of estimates converges. Prior to the first iteration, missing values for all analytes are initially filled in with the mean of the non-missing values for each analyte, but the algorithm keeps track of where the missing values were originally located. For every iteration thereafter, RF targets analytes one at a time to improve the estimate of imputed values. For each analyte, the algorithm selects only those samples that did not originally have a missing value for that analyte. From this set of samples, an RF model is generated, which aims to predict the target analyte based on its relationships to all other analytes in the dataset. Once constructed, this model is used to fill in the missing values of the target analyte again. The new value tends to be more accurate on each successive iteration of the imputation algorithm. A cycle of updating each analyte once represents a single iteration, and iterations continue until results converge.

In general, RF results are the aggregate of numerous decision trees, which are constructed by repeatedly splitting the data into smaller and smaller groups, or "nodes." For each split, samples are separated into one node or the other in such a way as to minimize the variance, or spread, among samples in the newly created nodes. Each resulting node is then split using this same criterion and so on until a node is either homogeneous (no variance) or contains five or fewer samples (the default minimum sample size criteria), at which point splitting stops along that branch. Because the algorithm aims to minimize within-node variance, nodes tend to become more homogeneous the further down the branches of the tree they are located. The samples that comprise the end nodes where the splitting has stopped are termed "leaves."

More specifically, each node in a decision tree has an associated decision rule defining how to divide the samples in the node into two new groups. The decision may be based on any predictor variable in the model. In the RF imputation algorithm, available predictors at any step of the analysis are all analytes except for the target analyte. The predictor variable to split on, and the particular value of that variable to split at, are chosen to minimize the variance in the resulting nodes. For example, if calcium was the target analyte, a splitting rule at some node might end up being that samples with magnesium values less than 100 mg/L go in one new node, and those with values of magnesium greater than or equal to 100 mg/L are assigned to the other.

Once all splitting rules have been defined, the RF can be used to predict response values for samples that were missing values for the target analyte. This is done by running these samples through each tree by following its decision rules, until each sample is assigned to a leaf. The mean response of the other samples in that leaf is then assigned as the new imputed value. The RF consists of many decision trees, and each is used to impute new values, and then the results are averaged to get a final value for each sample in that iteration.

Each iteration cycles through every analyte with missing values in this manner. After all analytes have had new imputed values assigned to them, the current iteration is then compared to the previous iteration, and the difference between their imputed values is measured. Another iteration is started if this difference is less than the difference measured from the previous iteration's comparison to its predecessor. If the difference is larger, the RF process is stopped and the current iteration becomes the output from the RF process. This output is used as the input in the next process, the generation of an RF dissimilarity matrix.

The imputation process was implemented using the *missForest* package in R (Stekhoven & Buehlmann 2012, Stekhoven 2013).

4.2 Random Forests Dissimilarity

When used to generate a similarity matrix instead of a regression model to predict values for imputation, RF compares the input data to a scrambled (random/synthetic) version of those data. This is necessary when no training dataset is available (i.e. unsupervised RF). Scrambling the data removes any relationship or correlation between concentration values within a sample. For example, it may be common in the original dataset for samples with high concentrations of arsenic to also have high concentrations of uranium; this correlation would not exist in the scrambled dataset. The algorithm now has two versions of the input data, one with the relationships between measured variables intact (the original data) and one without these relationships (the random/synthetic data). The original and synthetic data are combined, with each sample labeled according to whether it is an original or synthetic sample. After combining these two datasets, the RF algorithm builds trees that separate them back out. In doing so, the RF algorithm effectively learns about the relationships in the original data simply by contrasting it to the synthetic data. It should be emphasized that, like the imputed values descried above, the synthetic data used in unsupervised RF learning allow the algorithm to proceed technically, but they are not retained or analyzed further after the algorithm is complete. It is not uncommon for machine-learning algorithms to form fictitious data to help find relationships among the real data.

The RF unsupervised learning algorithm aggregates results from a large group of decision trees. The trees are similar to the regression trees described above (section 4.1), but in this case aim to distinguish among classes (real vs. synthetic) rather than predict values of an analyte. In classification problems, there is no variance to minimize, but RF uses the analogous criterion of attempting to maximize the homogeneity within each group produced by a split at a node. That is, each split in the RF unsupervised learning algorithm attempts to put mostly real observations in one group and mostly synthetic observations in the other group.

In RF unsupervised learning, the critical output of each decision tree is the similarity matrix, which is defined based on a simple rule: two samples are similar if they are classified in the same leaf of the tree, and different if they fall in different leaves. Each tree creates a similarity matrix crossing all samples with each other, and fills in a value of 1 for every pair that are similar, and a 0 for all dissimilar pairs. After all trees have been constructed, their similarity matrices are averaged together to get a composite similarity score between each pair of samples. Note that proximities are only calculated for the real samples, not the synthetic portion of the data used to help construct the RF.

For the baseline dataset fifty forests are included in this simulation, each with 1,000 trees. This results in a total of 50,000 trees. A reason not to simply run 50,000 trees in one forest is that each time a new forest is created, the input data is scrambled anew, lessening the chance that any one version of the scrambled data will impact the overall results.

The RF approach is unique in that it imposes no restrictions on the structure, distribution, or covariance of the data to be clustered, or the scale differences between variables or observations and, hence, offers a powerful and flexible means for identifying natural groupings in complex datasets.

The approach was implemented using the *randomForest* package in R (Liaw & Wiener 2002, Liaw *et al.* 2015).

4.3 PAM Clustering

The similarity matrix produced by the RF process is used as input to the PAM clustering algorithm. Based on the similarity values stored in the matrix, samples are clustered into ‘like’ groups. The number of groups, for this purpose, is pre-specified as two (one for background and the other for non-background samples). The PAM attempts to locate “medoids”—central values for a group of multivariate data points—around which to define the background and non-background clusters. For any such pair of medoids, samples are defined as “background” if they are closer to the background medoid than to the non-background medoid, and vice versa. The PAM algorithm works by searching for the pair that maximizes the overall similarity between sample points and the medoid of their assigned cluster. This process effectively separates the samples into the most distinct clusters possible. The approach was used to cluster samples into two groups for each of the eight stratigraphic units. A more in-depth description can be found in Appendix A.

4.4 Cluster Review

The expert review of the clustering results was performed on a well-by-well and sample-by-sample basis. Because some samples border on one cluster or the other, a review of sample classification was performed by subject matter experts with expertise in the Facility, its history, use and geology. This was also a method to verify the initial PAM clustering results.

Several actions were possible based on this review:

1) An entire well could be moved from background to non-background if it was onsite and in a known impacted area based on site history and events.
2) Samples from a well could be moved from baseline to non-baseline based on specific dates related to the history of the site and the spatial location of the well.
3) An entire well could be moved from non-baseline to baseline if it was known to be in an unimpacted area off-site (such as reclaimed areas northeast of the site) or upgradient of the site with potential impacts to the groundwater quality of the site from other sources.
4) Samples from a well could be moved from non-baseline to baseline based on review of the data, typically wells with samples that bounce around from cluster to cluster throughout the time period of record.

In total, this review process results in only 2.7% of 24,584 samples being switched from their PAM-assigned group. The result is a groundwater baseline dataset representing baseline conditions for the Facility.

4.5 Baseline Groundwater Dataset Finalization

Because one of the objectives of the current investigation is to calculate BSLs for a practical number of stratigraphic units, another review of the data across stratigraphic layers is done. Only the baseline data are used in this review in order to identify which units share similar baseline conditions and could therefore be combined for calculations of BSLs. A visual comparison of stratigraphic units (Figures 12-17) was performed, resulting in the four coal-related units (Rosebud Overburden, Rosebud Coal, Interburden, and McKay Coal) being combined into a single unit termed “Coal-Related.” Other units were left separate, resulting in five final units: Alluvium, Spoils, Clinker, Coal-Related, and SubMcKay. Figure 19 shows the spatial distribution of baseline wells across these final groundwater units, as well as the four baseline surface water sites. The Clinker baseline dataset is spatially limited in comparison to the other units (Figure 21). This is because the Clinker is so well drained that it does not contain much water. Overall the Clinker stratigraphic layer represents a smaller portion of the groundwater dataset. The spatial extent of the Alluvium, Spoils, Coal-Related, and SubMcKay baseline datasets is shown if Figures 20, 22, 23, and 24 respectively. Note that the sub-McKay may contain water in different positions in the depositional sequence. Sub-McKay wells are typically completed in the first water bearing interval below the McKay Coal. Where the first groundwater below the McKay Coal is encountered may vary over short lateral distances.

A comparison of this dataset to previous BSL datasets (Exponent 2011, Arcadis 2007, Maxim 2004) is provided in Table 6. The statistical summary of the current investigation dataset is provided in Table 5 and is comparable to Tables 2-5 in the previous Exponent investigation (2001), Table 5 in the previous Arcadis investigation (2007), and Table 3 in the previous Maxim investigation (2004).

Tables in Appendix H summarize the data in this final dataset as well as the data not used for BSL determination.

5.0 Baseline Screening Levels

In the current investigation, BSLs are represented statistically as the 95th upper confidence bound on the 90th percentile of concentrations for an analyte as observed in the baseline dataset. This is often referred to as a 95/90 upper tolerance limit (UTL). The methodology to do so combined bootstrapping with Gehan ranking and was implemented in R. Note that previous BSLs were calculated based on a 95/95 UTL (Arcadis 2007).

Bootstrapping (Efron 1993) works by drawing sets of values with replacement from observed samples many times, creating a simulation from the empirical data. Each realization of the sampling procedure provides an estimate of the 90th percentile. Each estimate is different, which creates a distribution for the 90th percentile. The 95th percentile of this distribution is interpreted as the 95th upper confidence bound of the 90th percentile.

Gehan ranking is a method used to account for censored data, such as detection limits. It is commonly used when performing nonparametric significance tests, such as the Wilcoxon rank sum test (Gehan 1965, Gilbert 1987, Helsel 2005, Martinez & Naranjo 2010, USEPA 2013), but its applications are much broader (e.g.. trend detection in water quality data, regression analysis, survival analysis). Gehan ranking treats non-detects as potentially representing any value less than the reported detection limit. The true value is unknown, but it has a maximum limit. All values (detects and non-detects) are ordered (i.e. ranked) lowest to highest based on their reported values and detection status. For each value, a new rank is determined by averaging all possible ranks the value could have. For example, a non-detect value may be originally ranked higher than a detect value, but the true value could be less, so this results in multiple possible ranks that are then averaged.

As an example of this approach, suppose there are four sample results with values [10, 20, 30, 40], and suppose the first and third sample results are non-detects [<10, 20, <30, 40]. Gehan ranking assigns the following ranks [1.5, 2.5, 2, 4]. That is, the first sample (<10) might be the lowest value sample, or the second lowest value (because of the non-detect at <30). The second sample (20) could occupy the 2nd or 3rd ranking position, hence its average rank is 2.5. The third sample (< 30) could occupy one of the first three ranking positions, hence its average rank is 2. Percentiles can then be computed based on the Gehan ranks of the data values.

To estimate a 95/90 UTL, the dataset is resampled with replacement, the samples are reordered according to the Gehan ranking scheme, and the 90th percentile is calculated. Each bootstrap realization provides a different estimate of the 90th percentile, which creates a distribution of the 90th percentile. The 95th percentile of this simulated distribution of 90th percentiles is the 95/90 UTL, which is used as the BSL. A more detailed example is shown in Appendix G.

The estimation of 95/90 UTLs using parametric methods depends heavily on the underlying distributional assumptions, and deviations from those assumptions can lead to poor 95/90 UTL estimates. This combined non-parametric approach makes no distributional assumptions and addresses non-detects with multiple detection limits easily. However, because of the relative novelty of applying Gehan ranking to UTLs, an additional, more common method of addressing non-detects was also applied for comparison purposes. This comparison is presented in Tables 8 and 10 alongside the Gehan-ranking based UTL estimates from this investigation. The alternate methods were a 95/90 UTL estimation using the reported detection limits (DL; reported non-detect values were not changed) and an estimation using half the reported detection limit (1/2 DL).

Tables 8 and 10 also show Circular DEQ-7 Human Health Standards for comparison.

5.1 Calculation of BSLs

For each analyte in each of the five stratigraphic units and in surface water, 90th percentiles and BSLs (95/90 UTLs) are computed, unless there are insufficient data to support BSL estimation. In cases in which where there were less than 10 samples, no BSL was calculated. It is common in environmental statistics to require sufficient useful data to perform reliable statistical analysis. In principle a mean and standard deviation can be estimated from two data points, but the estimation is not likely to be statistically reliable in the sense that a different two data points could provide very different estimates. Some consideration for the calculation of BSLs was given to the number of data points that might be needed. A decision was made that BSLs should not be presented if there are less than 10 data points in representing an analyte and/or stratigraphic unit. Table 4 summarizes the analytes across the stratigraphic units that have enough data points.

A large proportion of non-detects in a dataset can also make statistical analysis unreliable. For estimation of the 90th percentile, constraining the frequency of non-detects in the upper part of the data distribution is reasonable. Instead of applying additional rules that limit the number of BSL calculations performed, the decision was made to calculate BSLs for all remaining analyte and stratigraphic unit combinations (of size at least 10) and then flag certain 95/90 UTLs values according to the impact of the non-detects on the estimation.

For many of the datasets the non-detects are in the lower part of the data distribution and, hence, have no impact on estimation of the 95/90 UTL. In some cases, there are many non-detects, some of which appear in the upper part of the distribution, or even include the maximum reported value. This leads to three categories of calculated BSL values:

1. BSLs that are not impacted by non-detects. These BSLs are not flagged, and are considered the most reliable BSL estimates;
2. BSLs that are impacted by large valued non-detects; identified when the BSL is less than the 90th percentile of the reported values (i.e., when not adjusting for NDs). These BSLs are flagged as less reliable estimates; and,
3. BSLs that are the maximum reported value and this value is a non-detect. These BSLs are flagged as less reliable estimates.

These categories are identified in the BSLs comparison table (Tables 8 and 10) and in Tables 5 and 9. No shading applies to Category 1. Blue shading applies to Category 2. Orange shading applies to Category 3.

Note that Tables 5, 8, and 9 also flag estimated 90th percentiles (in yellow) that might be less reliable. The associated BSLs are also flagged. In these cases the top 20% of reported values are not all detects. This means some non-detects impact the BSL estimation. In some cases all of the data for a specific analyte and/or stratigraphic unit may have been non-detects.

5.2 Results

Table 1 lists the analytes and stratigraphic units for which there are sufficient data to estimate groundwater BSL calculations. A summary of the BSLs and the data used for their estimation is presented in Tables 5 and 7 for groundwater and Table 9 for surface water. Respective comparisons of BSL results from different statistical methods are presented in Tables 8 and 10 for groundwater and surface water. The two alternative methods differ only in how non-detects are handled. For these methods a 95/90 UTL is calculated using bootstrapping, but Gehan ranking is not applied. One uses the reported detection limit (DL) and the other uses half the reported detection limit (1/2 DL). These same tables also present MDEQ Circular DEQ-7 standards for human health (MDEQ 2012) for comparison.

The recommended groundwater BSLs are presented in Table 7. Empty cells indicate that no BSL was calculated because there was not enough available data. There are options for considering a different statistic to represent the upper end of baseline conditions when the BSL is not calculated or is unreliable. These options include the estimated 90th percentile and the maximum detected concentration reported in the baseline dataset. Note that the 90th percentile estimate is based on the raw data with no further manipulation of non-detects, and is sometimes greater than the maximum reported detected value because the greatest reported values are non-detects.

Highlighting is used to separate those combinations that clearly support groundwater BSL calculations from those for which the non-detects have an impact and thus result in less reliable estimations. If there is no shading in the cell, then the estimate is considered reliable (Category 1 in Section 5.1). The orange and blue highlighted results are considered less reliable. The estimated BSLs in these cases are impacted by non-detects. In effect the BSL might represent a detection limit, rather than concentration data.

For highlighted BSL results, this indicates one of two unique cases for which the given BSL value requires careful consideration. In the blue shaded cases (Category 2 in the Section 5.1) the estimated BSLs are less than the 90th percentile (when the data are evaluated without adjustment for non-detects). An example of this is lead within SubMcKay (Table 5). There is a large number of relatively large-valued non-detects within the dataset, and the maximum non-detect is nearly as great as the maximum detect value. Large-valued non-detects can affect estimation of the 90th percentile, but the conditions of the effect depend on where the non-detects are in the original ordering of the data.

In the orange shaded cases (Category 3 in Section 5.1), the BSLs correspond to the largest reported value and that value is also a non-detect. This generally occurs when the maximum value is a non-detect and the detect frequency is low. However, this condition is hard to predict because it also depends on the general occurrence of non-detect values within the detected values. An example is titanium within the Alluvium unit. Table 5 shows the largest value is 0.1 and it is a non-detect. Even with Gehan ranking, this value gets chosen enough times in the bootstrapping simulations as the 90th percentile that it becomes the BSL. The large number of non-detects relative to detects in this titanium dataset makes interpretation of the BSL and the 90th percentile difficult.

Yellow shading is used for the 90th percentile estimates in Tables 5, 8, 9, and 10 when there are any non-detects in the top 20% of the reported values. In general, colored shading implies estimates that are affected by non-detects. If the effect is at the 90th percentile then that effect also carries into the BSL, which may have orange or blue shading indicating a more specific non-detect effect. In cases where the 90th percentile is shaded yellow, but no additional shading is present in the BSL column, careful interpretation of the BSL value is still recommended. For example, nickel in Alluvium (Table 5) has a 90th percentile shaded in yellow, but no shading for the BSL value. This still indicates that the BSL might be representative of detection limits rather than actual concentrations. In fact, this nickel dataset has a detect frequency of only 13.2 percent. Most of these data represent detection limits. One possible interpretation is that baseline conditions are unknown and the BSL shown should not be applied.

However, because the BSLs will be applied only if they exceed human health standards (see Tables 8 and 10 for these), shaded BSLs are only of concern where they exceed these standards. In the case of the nickel example above the calculated BSL value is the same as the human health standard of 0.1 mg/L. So while the shown BSL for alluvial nickel should be interpreted carefully, it has little bearing on applied screening levels. There are some groundwater BSLs that exceed the human health standards and should also be interpreted carefully due to potential effects by non-detect values included in the BSL estimation. These are:

- Tin in Alluvium,
- Beryllium in Spoils and Coal,
- Lead in Spoils,
- Mercury in Spoils,
- Nitrite in Coal,
- Thallium in Coal and SubMcKay

The alluvial tin BSL is the maximum non-detected value, as is the BSL for beryllium and nitrite in Coal and mercury in Spoils. Beryllium concentrations in Spoils are all non-detects except one. Thallium has no detects in either Coal or SubMcKay. The amount of data in these cases that are non-detects suggests the BSL is primarily calculated based on detection limits and not measured concentrations. The BSL for lead in Spoils is based on a 30% detection frequency, the highest by far of any of the analytes listed above. However, the fact that the BSL (detect status is a consideration) is less than the 90th percentile (detect status is not a consideration) suggests it too is influenced by non-detects and should be interpreted with caution.

In surface water, only mercury exceeds the human health standard and has a BSL that may be affected by non-detect values. In this case, the BSL is a non-detect value and the real baseline concentration is lower.

References

Anning, David W., Paul, Angela P., McKinney, Tim S., Huntington, Jena M., Bexfield, Laura M., and Thiros, Susan A. (2012). Predicted nitrate and arsenic concentrations in basin-fill aquifers of the southwestern United States. U.S. Geologic Survey National Water-Quality Assessment. Scientific Investigations Report 2012-5065.

Arcadis, (May 22, 2007). Data Analysis and Statistical Evaluation of Unimpacted Groundwater Quality.

Breiman, Leo (2001). Random Forests. Machine-learning, 45(1):5–32.

Roberts, S.B., Wilde, E.M., Rossi, G.S., Blake, Dorsey, Bader, L.R., Ellis, M.S.,Stricker, G.D., Gunther, G.L., Ochs, A.M., Kinney, S.A., Schuenemeyer, J.H., and Power, H.C. (1999). Colstrip Coalfield, Powder River Basin, Montana: Geology, Coal Quality, and Coal Resources *in* U.S. Geological Survey Professional Paper 1625-A.

Cutler, Richard D., Edwards, Thomas C., Jr., Beard, Karen H., Cutler, Adele, Hess, Kyle T., Gibson, Jacob, and Lawler, Joshua J. (2007). Random forests for classification in ecology. Ecology, 88(11):2783-2792.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. Bioinformatics, 7(3).

Efron, B., and Tibshirani, R. (1993). An Introduction to the Bootstrap, Chapman and Hall, CRC Monographs on Statistics and Applied Probability.

Evans, S. Jeffrey and Cushman, Samuel A. (2009). Gradient modeling of conifer species using random forests. Landscape Ecology, 24:673-683.

Exponent, (April 18, 2011). Data Analysis and Statistical Evaluation of Unimpacted Groundwater Quality, Units 3 and 4 Effluent Holding Pond Area, Colstrip Steam Electric Station, Colstrip, Montana.

FordCanty & Associates, Inc. (2015). Cleanup Criteria and Risk Assessment Work Plan. Wastewater Facilities Comprising the Closed-Loop System Plant Site Area, Colstrip Steam Electric Station, Colstrip, Montana. October.

Gehan, Edmund A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika, 52(1-2):203–23.

Gilbert, Richard, O. (1987). Statistical Methods for Environmental Pollution Monitoring. Van Norstrand Reinhold, New York. 320 pp.

Gower, J.C. and Legendre, P. (1986). Metric and Euclidean Properties of Dissimilarity Coefficients. Journal of Classification, 3:5-48.

Helsel, Dennis R. (2005). Nondetects and Data Analysis: Statistics for Censored Environmental Data, John Wiley and Sons, Inc., Hoboken, NJ.

Ho, H.C., Knudby, A., Sirovyak, P., Xu, Y., Hodul, M., and Henderson, S.B. (2014). Mapping maximum urban air temperature on hot summer days. Remote Sensing of Environment, 154:38-45.

Hydrometrics (2015). Evaluation of 2014 Hydrologic Monitoring Data from Colstrip Units 1 through 4 Process Pond System Colstrip Steam Electric Station, Colstrip, Montana.

Hydrometrics (2015). PPL Montana, LLC Colstrip Steam Electric Station Administrative Order on Consent Plant Site Report. July.

Hydrometrics (2014). Evaluation of 2013 Hydrologic Monitoring Data from Colstrip Units 1 Through 4 Process Pond System, Colstrip Steam Electric Station, Colstrip, Montana. April.

Hydrometrics (2013). PPL Montana, LLC Colstrip Steam Electric Station Administrative Order on Consent Units 1&2 Stage I and II Evaporation Ponds Site Report. May.

Hydrometrics (2013). PPL Montana, LLC Colstrip Steam Electric Station Administrative Order on Consent Units 3&4 Effluent Holding Pond (EHP) Site Report. October.

Kaufman, L. and Rousseeuw, Peter J. (1987). Clustering by Means of Medoids. Statistical Data Analysis Based on the L1-Norm and Related Methods, 405–16. North-Holland.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2(3), 18–22.

Liaw, A., Wiener, M., Breiman, L. and Cutler, A. (2015). Breiman and Cutler's Random Forests for Classification and Regression (Package "randomForest") (version 4.6-12). https://www.stat.berkeley.edu/~breiman/RandomForests/.

Martinez, R. L. M. C. and Naranjo, J. D. (2010). A pretest for choosing between logrank and wilcoxon test in the two-sample problem, International Journal of Statistics, LXVII (2): 111-125.

Mascaro, J., Asner, Gregory P., Knapp, David E., Kennedy-Bowdoin, Ty, Martin, Roberta E., Anderson, Christopher, Higgins, Mark, and Chadwick, K. Dana. (2014). A tale of two "forests": Random forests machine-learning aids tropical forest carbon mapping. PLoS One 9(1).

Maxim. (2004). Preliminary Conceptual Hydrogeologic Model, Stage I and I1 Evaporation Ponds and Plant Site Areas, CSES-Colstrip, Montana." Maxim Technologies Inc.

MDEQ. (2012). Circular DEQ-7: Montana Numeric Water Quality Standards. MDEQ, Water Quality Planning Bureau, Water Quality Standards Section.

MDEQ/PPL Montana (2012). Administrative Order on Consent Regarding Impacts Related to Wastewater Facilities Comprising the Closed-Loop System at Colstrip Steam Electric Station, Colstrip Montana.

Mellor, A., Haywood, A., Stone, C., and Jones, S. (2013). The Performance of Random Forests in an Operational Setting for Large Area Sclerophyll Forest Classification. Remote Sensing, 5: 2838-2856.

Microsoft. (2011). Microsoft Excel [computer software]. Redmond, Washington: Microsoft.

Montana Bureau of Mines and Geology (2015). Groundwater Information Center Mapping Database, online at http://data.mbmg.mtech.edu/mapper/mapper.asp?view=Wells&.

Neptune and Company, Inc. (2015). Work Plan for Development of Updated Background Screening Levels.

Prasad, Anantha M., Iverson, Louis R., and Liaw, Andy. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems, 9:181-199.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Roberts, S.B., Wilde, E.M., Rossi, G.S., Blake, Dorsey, Bader, L.R., Ellis, M.S.,Stricker, G.D., Gunther, G.L., Ochs, A.M., Kinney, S.A., Schuenemeyer, J.H., and Power, H.C. (1999). Colstrip Coalfield, Powder River Basin, Montana: Geology, Coal Quality, and Coal Resources. U.S. Geological Survey Professional Paper 1625-A.

Rodriguez-Galiano, Victor, Mendes, Maria Paula, Garcia-Soldado, Maria Jose, Chica-Olma, Maria, and Ribeiro, Luis. (2014). Predictive modeling of groundwater nitrate pollution using Random forests and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Souther Spain). Science of the Total Environment, 476-477:189-206.

Shah, Anoop D., Bartlett, Jonathan W., Carpenter, James, Nicholas, Owen, and Hemingway, Harry. (2014). Comparison of random forests and parametric imputation models for imputing missing data using MICE: A CALIBER study. American Journal of Epidemiology, 179(6):764-774.

Shi, T., Sligson, D., Belldegrun, A.S., Palotie, A. and Horvath, S. (2005). Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma. Modern Pathology, 18:547–57.

Stekhoven D. J., & Buehlmann, P. (2012). MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112-118.

Stekhoven, D. J. (2013). missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4.

USEPA (2013). Pro UCL Version 5.0.00 Technical Guide.

Wolfe, F., Clauw, D.J., Fitzcharles, M.A., Goldenberg, D.L., Katz, R.S., Mease P., Russell, A.S., Russell, I.J., Windfield, J.B., and Yunus, M.B. (2010). The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. Arthritis Care Research (Hoboken), 62(5):600-610.