Natural Resources Research (© 2021)
https://doi.org/10.1007/s11053-020-09801-5

Original Paper

Estimating Stable Measured Values and Detecting Anomalies in Groundwater Geochemistry Time Series Data Across the Athabasca Oil Sands Area, Canada

John G. Manchuk,[1,7] Jean S. Birks,[2] Cynthia N. McClain,[3] Guy Bayegnak,[4] John J. Gibson,[5] and Clayton V. Deutsch[6]

Received 1 May 2020; accepted 24 December 2020

Regional groundwater monitoring in the Athabasca region of Alberta, Canada, provides information on groundwater quality and geochemical changes over time, including data useful for evaluating potential impacts of industrial activity such as oil sands mining and in situ operations. Data collected from over 5000 wells from the 1950s to 2014, including 161 wells from government's monitoring network, were used to develop and apply bootstrap techniques for the detection of changes in groundwater geochemistry over time and at specific points in time. Increasing temporal anomalies were identified in Cl, TDS, B, and naphthenic acids in the McMurray formation across 2003 and 2008, while decreasing anomalies were found for SO_4. Temporal variance for 15 indicators was quantified for a smooth bootstrap approach to arrive at stable values representative of the most recent samples taken from wells in the study area. Stable values revealed sampling bias in the Devonian, Grand Rapids, Empress, Channel Beverly, and Muriel Lake formations suggesting expansion of sampling may be necessary. Although temporal anomalies were found in the McMurray formation, sampling bias was not identified. The entropy and relative magnitude of time series were evaluated to identify candidate wells for continued observations, which consist of wells with low measurements and low entropy that are near active industry lease boundaries. Temporal anomalies, stable values, and entropy were combined into type-well information to provide plots for visual inspection and interpretation. Stable values are useful for regional mapping, for detecting future changes and trends, and for identifying areas of interest warranting further investigation.

KEY WORDS: Geostatistics, Geochemistry, Bootstrap, Water monitoring, Water quality.

[1]Centre for Computational Geostatistics, 6-050 Natural Resources Engineering Facility, University of Alberta, Edmonton, AB T6G 2W2, Canada.
[2]InnoTech Alberta, 3608-33 St NW, Calgary, AB T2L 2A6, Canada.
[3]Alberta Environment and Parks, 3535 Research Road NW, Calgary, AB T2L 2K8, Canada.
[4]Alberta Environment and Parks, 7th Floor, Oxbridge Place, 9820-106 Street, Edmonton, AB T5K 2J6, Canada.
[5]InnoTech Alberta, 3-4476 Markham Street, Victoria, BC V8Z 7X8, Canada.
[6]Centre for Computational Geostatistics, 6-247 Donadeo Innovation Centre for Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada.
[7]To whom correspondence should be addressed; e-mail: jmanchuk@ualberta.ca

Published online: 18 January 2021

INTRODUCTION

Analysis of temporal geochemistry samples taken from hydrostratigraphic units surrounding industrial activity is an important measure for detecting changes in the natural composition of groundwater. Commonly, due to lag between industrialization and development of regional monitoring policy, it can be difficult to characterize predevelopment conditions representative of the natural composition of groundwater. In such reactive scenarios, it is often only practical to estimate the water composition at the present time, although such information remains useful for detecting future changes. This is the case for the Athabasca oil sands area where activity related to bitumen mining from the McMurray Formation began in the 1930s (Parker and Tingley 1980), whereas the first systematic groundwater sampling initiative was not conducted until the 1950s (Akena and Christian 1981). While substantial industrial activity did not begin until the late 1960s, this preceded establishment in 1976 of the first systematic regional water quality monitoring program by the Alberta Oil Sands Environmental Research Program (AOSERP). Surficial mining operations expanded further from 2000 to 2013, and in situ production surpassed mined oil sands production in 2012. Implementation and maintenance of such a monitoring program is important because there are potential ecological and human health concerns, for example due to polycyclic aromatic hydrocarbons (PAH) and arsenic releases associated with surface and in situ industrial activity (Kelly et al. 2009; Javed and Siddique 2016). To address this, a Groundwater Management Framework (GWMF) and supporting strategic plans were released by Alberta Environment and Parks as a part of the Lower Athabasca Regional Plan to manage the cumulative effects of activities in the region (AEP 2012). As well, the industry funded (up to $50 M/year) joint Federal Provincial Oil Sands Monitoring Program is responsible for design and implementation of environmental monitoring to track baseline conditions and environmental impacts in the region (Government of Alberta 2017), including groundwater monitoring. This Oil Sands Monitoring Program informs development of policy and plans through integrated reporting and analysis of environmental condition.

A substantial confounding factor limiting ability to detect changes in water quality is the presence of gradients in geochemical composition that exist due to factors unrelated to anthropogenic activity. This pertains to naturally occurring elements that directly impact an aquifer. Detection of compounds or those without natural sources in the region are likely related to industrial or agricultural activity; however, this is challenging in the oil sands region because many of the solutes and organics associated with oil sands process water are also naturally present in groundwater. Irrespective of the cause of a detected change, further analysis would be necessary to identify a source as natural or industrial. A statistical technique to detect changes is a critical component to a temporal water quality monitoring program. Research into change-point detection has led to several techniques related to water quality time series (Ba and McKenna 2015). Several categories of change-point detection methods exist, with frequentist methods being the most popular due to their simplicity and success rate (Ghosh and Sen 1991). The sequential probability ratio test of Wald (1945), one and two-sided sequential cumulative schemes of Page (1954) and Lorden (1971), and likelihood ratio tests (Siegmund and Venkatraman 1995) are examples of frequentist methods for change-point detection. These and associated methods have been compared by Roberts (1966) and continue to be compared in more recent research such as Ba and McKenna (2015) and Breitenberger et al. (2018). A variety of other categories exist that utilize machine learning strategies such as support vector machines (SVM) and artificial neural networks (ANN) among others. Lauzon and Lence (2010) use ANNs to detect shifts in time series data, where a shift is analogous to a change-point. SVMs were utilized by Camci (2010) for the change-point detection problem for increases or decrease in the mean or variance of a time series.

Considering that a change-point is a point in time where a statistic has increased or decreased, it is also likely that other classical techniques may be viable. If such a point is assumed known, the problem is analogous to hypothesis testing for a difference in the mean, or a difference in the variance of a sample distribution. The classical Student-t test is relevant for certain problems (Johnson and Bhattacharyya 1996). A variety of other parametric tests are available for comparing means of two possibly correlated distributions. Such tests rely on assumptions about the shape of the generating distributions and stationarity, which may not be applicable to water quality monitoring samples, in particular, those with temporal trends, noise, and censoring.

Nonparametric methods such as bootstrap techniques (Efron 1979; Hardle et al. 2003) are applied to a number of problems in water quality monitoring. An autoregressive moving average (ARMA) bootstrap approach was used by Nordgaard and Grimvall (2006) for conducting Mann–Kendall tests for the presence of trends (Mann 1945; Kendall 1975). Anttila et al. (2012) use moving block bootstrap simulations to evaluate sampling errors for quantifying temporal representativeness of water quality data. Resampled bootstrapping was used by Keum and Kaluarachchi (2015) to estimate confidence intervals of incremental dissolved solids yields based on a SPARROW transport model. Other types of nonparametric techniques have been applied to groundwater monitoring as well (Helsel and Hirsch 2002); for example, generalized least squares and the Tukey–Kramer method for mean comparison were used by Rodvang et al. (2004) to assess changes in ground water quality in southern Alberta.

For the Athabasca Oil Sands Groundwater Monitoring Program (AOS-GWMP), water quality time series sampling has been relatively infrequent and inconsistent as compared to many examples found in the literature. For example, online monitoring networks can utilize in situ sensors to collect data anywhere from monthly down to daily or hourly frequency, yielding relatively dense time series for certain parameters such as turbidity, electrical conductivity, pH, temperature, ammonia-N, nitrate-N and phosphate (Iwanyshyn et al 2009). In contrast, time series from the AOS-GWMP are sampled every 4 months or longer and somewhat inconsistently, leading to extended periods of inactive sampling. Samples below detection limit as well as noise are concerns for analysis; hence, this work utilized nonparametric resampling bootstrap approaches to test for the significance of manually defined temporal change-points. The bootstrap-t approach was used to conduct a two-tailed t-test for the difference in time series before and after a potential change-point (DiCiccio and Efron 1996). Due to limited sample sizes, smoothing methods were also utilized (Polansky 2000) along with spatial data aggregation that is characteristic of block bootstrap techniques applied in spatial and temporal domains (Lahiri 1999; Buhlmann 2002).

In addition to testing for change-point significance, quantifying stable measured values at the present time is important for water quality monitoring projects. Noise observed in time series of water quality measurements can be related to short-term environmental events, seasonal variation, localized concentration gradients, ion distribution, and other natural processes that must be considered to provide a stable reference value. Sampling practices and laboratory equipment may also contribute to noise. Establishing groundwater composition at the present time provides a baseline for detecting changes or for use as initial conditions for numerical water quality modeling (Palmer 2001). Defining a spatial–temporal water quality model would provide insight into the movement of groundwater as well as give some capability to predict future changes; however, defining the necessary initial and boundary conditions and tuning such a model to match historical water quality measurements is a daunting task. Such a model would be relevant to identifying contaminant sources and also for planning future monitoring activity including the placement of new monitoring wells and revision of sampled variables at existing sites.

For the variables of interest, sample availability and a statistical analysis of the temporal distributions is performed to identify major time intervals of data collection and important intervals for detection of changes. A bootstrap-t approach (DiCiccio and Efron 1996) is used afterward to test for time-based anomalies or changes in the magnitude of key variables. The results from time anomaly detection suggest that a viable approach for estimating stable measured values of the variables is a smooth bootstrap approach with a variance derived from the time dimension. Sampling distributions are analyzed to assign sampling scores to each of the monitoring wells to identify areas considered interesting, that is, areas that are significantly different from the global distributions and that could benefit from additional monitoring.

MATERIALS AND METHODS

Dataset

The groundwater quality monitoring program data consist of spatial–temporal samples recorded from 1958 to 2014 covering a large portion of the Alberta oil sands resource, see Figure 1. Due to diverse hydrogeological conditions and development pressures, the region can be divided into three discrete areas: the Northern Athabasca Oil Sands (NAOS), Southern Athabasca Oil Sands (SAOS), and the Cold Lake-Beaver River (CLBR). The da-

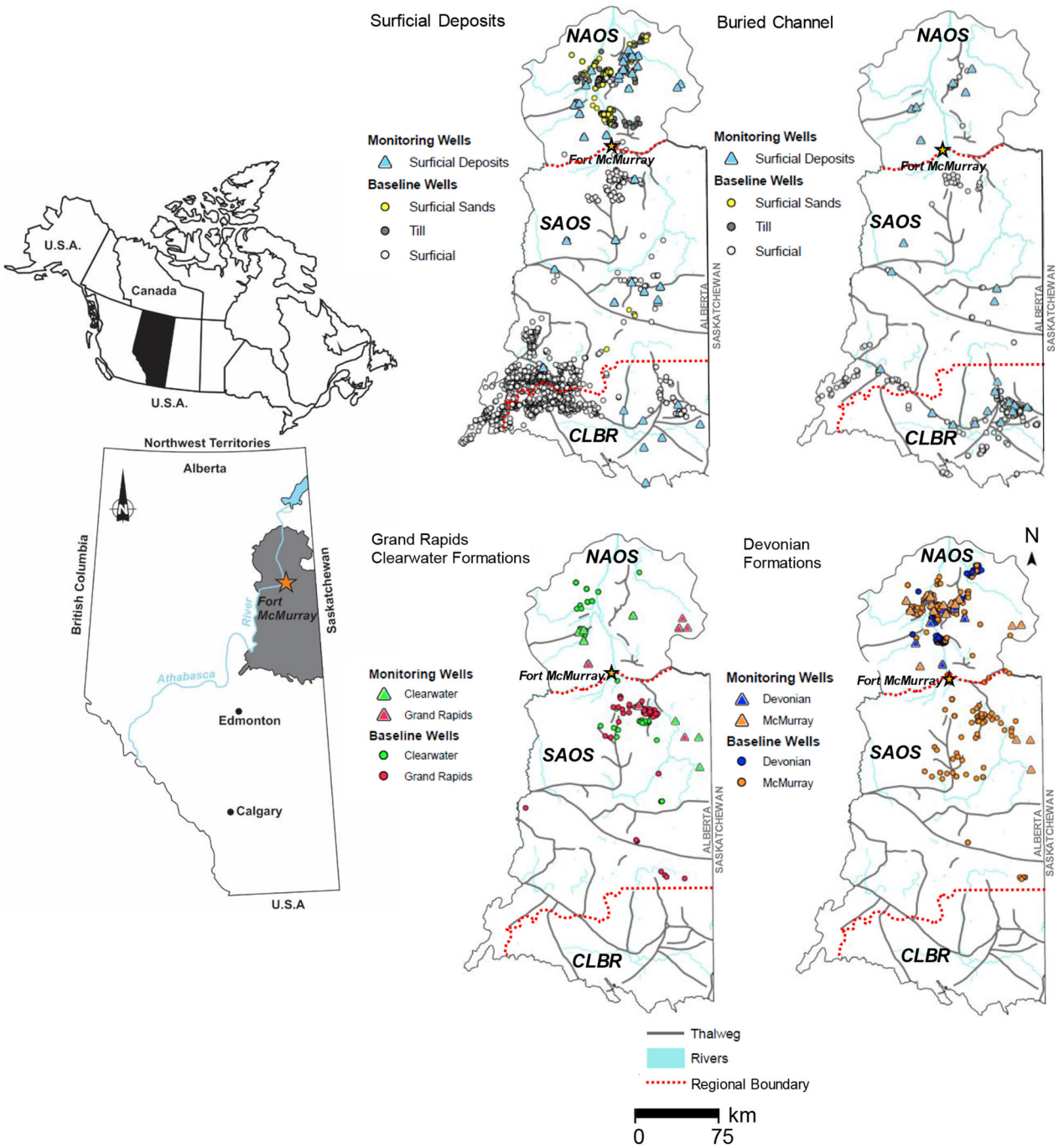

Figure 1. Distribution of oil sands monitoring wells (MW) (triangles) and baseline data (circles) for the main hydrostratigraphic units in the region. Surficial deposits includes all Quaternary and Neogene formations including undifferentiated overburden, while Channel wells show wells specifically identified as being in channel deposits. Image from Birks et al. (2019).

taset consists of samples from over 5000 wells that are categorized as industry or government wells, with the latter being monitored on behalf of Alberta Environment and Parks (AEP) by contractors. One hundred sixty-one wells are part of AEP's active monitoring network (Fig. 1). Water chemistry samples were obtained from wells drilled into various aquifers and groundwater sources. Bedrock forma-

tions such as the Devonian, Clearwater, and the McMurray were sampled as well as surficial sands and undifferentiated overburden material. Five hundred sixty-four different variables were sampled in the network, albeit inconsistently, and not necessarily for the purposes of the monitoring network as it pertains to industrial activity.

Data sources were provided by the Alberta Environmental Monitoring, Evaluation and Reporting Agency (AEMERA), AEP, and Alberta Energy Regulator (AER) or obtained from publically available datasets (e.g., Environmental Impact Assessments [EIA's]). Where necessary, data were obtained from the following additional sources: a compilation by AEMERA for the GWMF; a recent monitoring report done by Matrix Solutions Inc. (2015a, b); and data from EIA's available from AEP. Data requests can be submitted to AEP.GOWNinfo@gov.ab.ca. Ion balances were determined for each groundwater observation and samples with less than 10% error were kept in the dataset. A pH criterion of $pH > 6$ and $pH < 9.2$ was used as a quality control criterion to eliminate samples that were not representative of groundwater formation conditions.

Indicators considered for this paper consist of those of relevance for human or ecological health, associated with natural or industrial contamination, or which have been assigned by the GWMF critical upper thresholds, called interim triggers that indicate water quality concerns for investigation (AEP 2012; Matrix 2015a, b; Birks et al. 2019). Fourteen of the indicators were assigned interim triggers by AEP such as carcinogenic organic components of bitumen or process waters from oil sands mining and mobilized arsenic from in situ operations. Indicators used for examples in this work, denoted the indicators of interest, include total dissolved solids (TDS), sodium (Na), chloride (Cl), sulfate (SO_4), arsenic (As), silica (Si), nitrogen (N), boron (B), naphthenic acids (NA), temperature (T), and dissolved and total organic carbon (DOC, TOC). Two variants of N were sampled: as dissolved nitrate (NO_3) and as ammonia (NH_3). Phenols were also assigned an interim trigger by AEP; however, their data quality is not of a sufficient standard for analysis. Two specific PAH indicators categorized with the highest proportion of samples above detection were pyrene (PY) and naphthalene (NAPH). Although these are not specific indicators in the GWMF, they are included in the analysis that follows. Sample availability of these indicators was assessed spatially and temporally. Availability is measured in terms of the number of wells that had a particular indicator sampled, and also in terms of the total number of samples available. The number of wells gives an indication as to the spatial coverage of an indicator, with more wells generally indicating better coverage. Spatial coverage is confounded by clustering and stratification from different formations. The total number of samples for a given indicator provides a general idea of the sample support for calculating associated statistics.

Groundwater quality samples are collected from monitoring wells that target different stratigraphic layers encountered in the subsurface that form hydrostratigraphic units. Different units may be independent if groundwater and/or surface water are separated by an impermeable barrier such as a shale layer, or they may communicate if such barriers do not exist or if zones with elevated permeability associated with faults or fractures are present. Examples of hydrostratigraphic units that are mined in the Athabasca region include the McMurray and Wabiskaw Formations; those developed via in situ methods include the McMurray and Wabiskaw formations; and units of interest as sources of non-saline groundwater include surficial deposits, buried channel deposits, and portions of the McMurray, Clearwater and Grand Rapids formations in the NAOS and SAOS regions as well as the Sand River, Ethel Lake, Bonnyville, Muriel Lake, and Empress formations of the CLBR area (Fig. 1).

Note that surficial sediments are unconsolidated glacial and pre-glacial gravels, fluvial sands, tills and lacustrine clays. These are also called surficial deposits, or drift and include surficial sands, undifferentiated overburden and channels. The locations of mapped thalwegs of major channels are shown in Figure 1. The thickness of surficial sediments generally decreases toward the north ranging from 10 to 30 m thick. Surficial sediments can exceed 100 m in deeply incised channel formations eroded into bedrock (for example, the Helina, Beverly, and Empress Channels). In the Cold Lake Beaver River region, multiple Quaternary aquifers composed of surficial sediments, alternating with aquitard units, have been defined (Andriashek 2003). These formations include Sand River, Ethel Lake, Bonnyville, Muriel Lake, and the Empress formation, which is the largest Quaternary aquifer in the CLBR region and is primarily composed of coarse fluvial sediments. Bedrock of Cretaceous age is comprised of a sedimentary sequence that dips to the southwest and

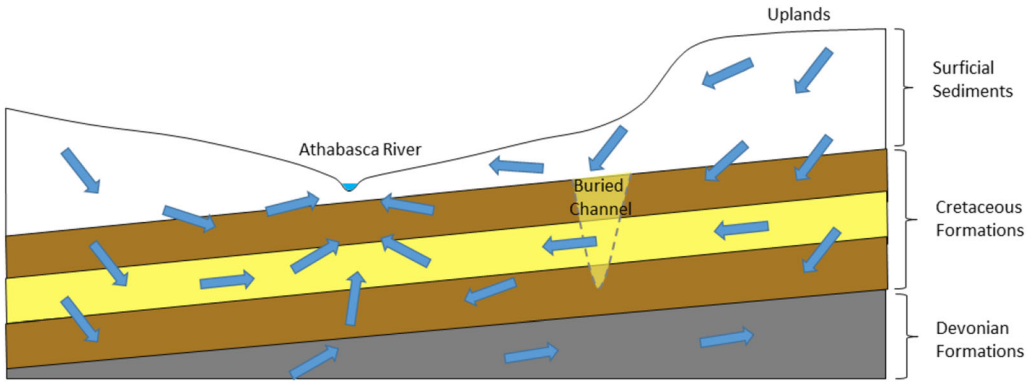

Figure 2. Schematic cross section (not to scale) showing general groundwater flow directions.

subcrops in the greater Fort McMurray area. Aquifers of the Manville Group are used as industrial, low-quality water sources and include the Grand Rapids, Clearwater, and McMurray formations. The McMurray formation is mined for bitumen deposits. Cretaceous bedrock is unconformably underlain by Devonian limestone and evaporites (e.g., halite and anhydrite) which subcrop along the valley walls of the Athabasca and Clearwater rivers and have formed dissolution and collapse features creating vertical connectivity for migration of high TDS waters. Figure 2 shows a schematic cross section with general groundwater flow directions. For more review of geology in the region, refer to Ranger and Gingras (2003). For more on groundwater, refer to Bachu et al. (1993).

Due to the areal extent of approximately 400 km north–south by 200 km fast–west (142,000 km^2) and depth of sampling ranging from near surface to roughly 600 m depth, a significant amount of heterogeneity is encountered adding to the complexity of analysis. Indicator concentrations may vary by several orders of magnitude across the area, for example, TDS ranges from 1 mg/L to over 400,000 mg/L (Fig. 3).

Availability was checked by hydrostratigraphic units as they define the aquifers that are geologically unique in terms of spatial extent, rock properties, water, and bitumen contents. However, it is possible that aquifers are connected either through direct contact in the depositional hierarchy, or through open faults and collapse features that are typical of the area. Availability of indicators by sample count is provided in Figure 4. Not all indicators are sampled from all units, and not all indicators are sampled equally within each unit. Units are presented

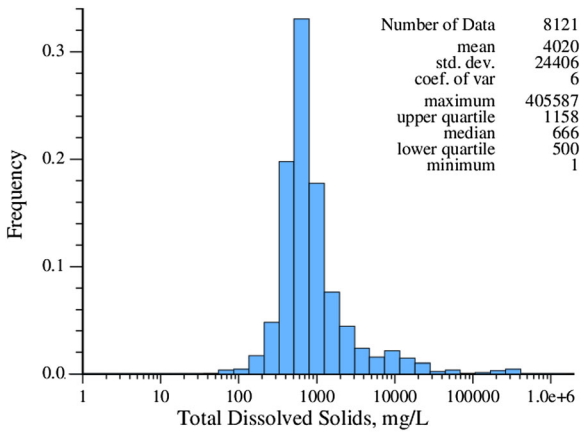

Figure 3. Histogram of TDS from the AOS-GWMP.

according to their geological succession, from the youngest to the oldest, younger units typically shallower than older units; however, due to the geological processes such as hiatus, tectonics, dissolution, and erosion for example, the relationship between depth and age may not always be respected. For example, the McMurray formation may occur near or at ground surface where it outcrops and is typically targeted for oil sands recovery to a depth of 75 m or greater where in situ operations are practical. Naphthalene and pyrene are above detection in a large number of units, and most are above detection in the surficial units, Clearwater, Devonian, and McMurray formations.

Temporal availability of the indicators was also assessed. Analysis quantified the number of samples that were available for each indicator and for any given sampling year, with time ranging from 1958 to 2014, see Figure 5. Prior to 1958, sampling was rare

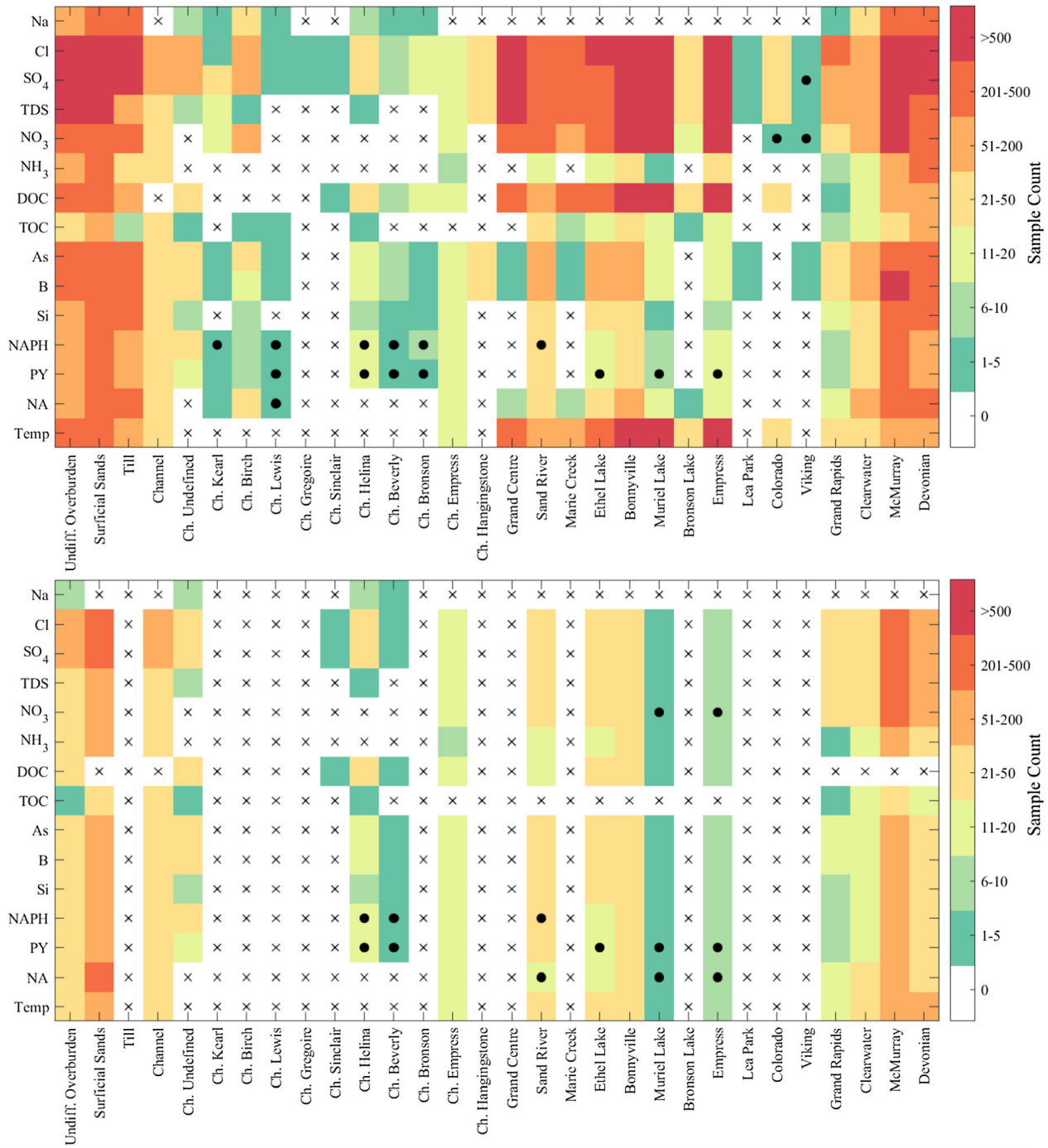

Figure 4. Availability of indicators of interest colored by the number of samples by hydrostratigraphic unit for all wells (top) and AEP's active monitoring wells (bottom). An x denotes no data; bullets indicate all samples were below detection.

and sporadic, rendering a very low confidence in the data. TDS, Cl, and SO_4 were measured consistently for a much longer time span than the other indicators of interest, while other indicators, including Na, NH_3, TOC, As, B, Si, NAPH, PY, and NA, were measured infrequently prior to 1990. Temporal

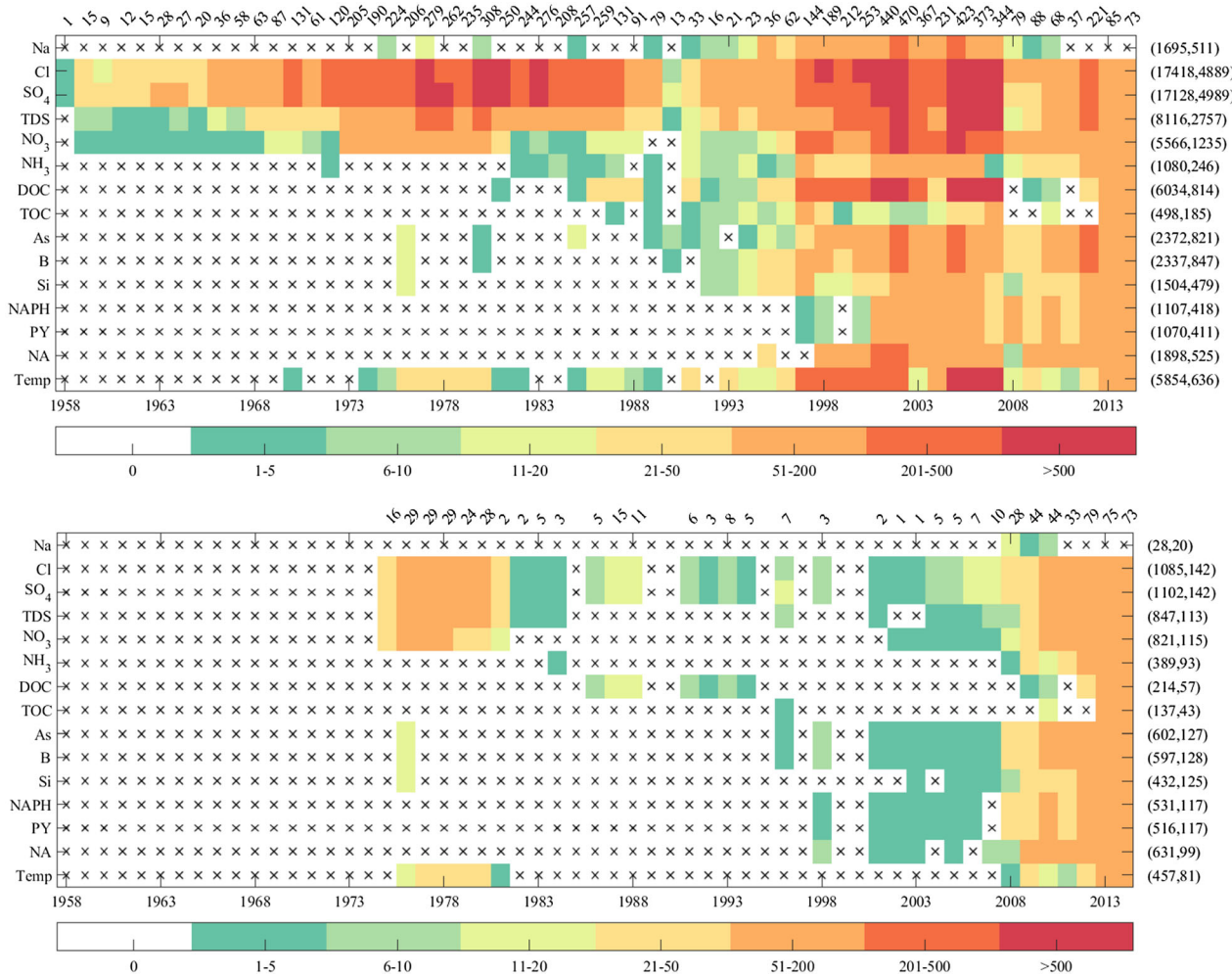

Figure 5. Timeline of indicators of interest from 1958 to 2014 for all wells (top) and AEP's active monitoring wells (bottom). An *x* indicates no samples. Numbers along the top indicate wells sampled in that year for all variables combined. Numbers at the right indicate (total samples, total wells) sampled for each indicator over the time span.

trends in sampling are observed in Figure 5 including: from 1997 to 2007 a period of increased sampling is observed compared to any other interval for all wells combined; sampling appears to be consistent from 2012 and on; for AEP's active monitoring wells, a significant gap in monitoring data exists from 1980 to 2008 and an increasing trend in sampling is clear from 2008 to 2014, suggesting progress toward expanding the network. Since 2010, the indicators that are the focus of this work were entirely sampled from AEP's active monitoring well network. Points of significant drops in sampling across all wells occurred from 1986 to 1990, in 2003, and in 2008. At least 10 out of 12 indicators have been sampled between 1992 and 2014.

Sample Variance

Understanding the quality of the sampling aided in method selection for time anomaly tests and establishing stable estimates of measurements. Consistency of sampling was evaluated using run lengths and variability was quantified using a nugget effect calculated in the time dimension. The occurrence of short time series was far greater than long time series in the sample data. The global distribution of time series lengths evaluated as run lengths with different break intervals was computed for the indicators of interest combined (see Figure 6). Break intervals are time gaps where the run terminates if a gap longer than the break is encountered. A 1-year break has little effect, while a 1-month break con-

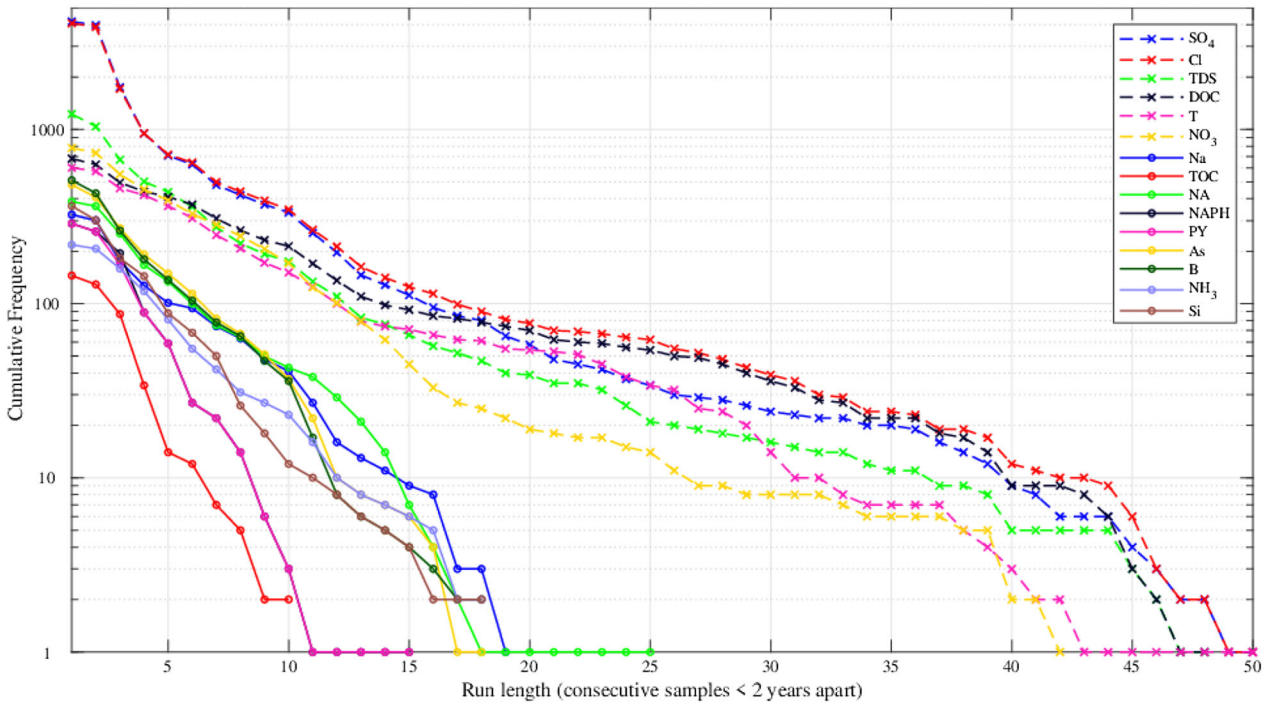

Figure 6. Runs lengths for the indicators of interest. Runs were broken if the time separating two consecutive samples was greater than 2 years.

siderably reduces the number of time series lengths that are over five to ten samples long. The average time interval between sampling is 6 months for approximately converting the run lengths into time. There are very few runs with more than 20 points in the data, where beyond this length there are no more than 40 occurrences for any given length.

Figure 6 shows the run length distributions for the indicators of interest with a break interval (time separating consecutive samples) of 2 years. Values are cumulative in nature in that the number of run lengths of size x is equal to the sum of all runs of size $y \geq x$. For example, for temperature, there are roughly 15 runs or time series with a length of at least 30 points, or for Si, there are two time series with run lengths of at least 16. Two groups of run-length distributions are observed in Figure 6 with long runs and short runs. Short runs coincide with indicators that were rarely sampled prior to 1990, while variables with long runs coincide with those sampled more regularly over the time span of the monitoring project including SO_4, Cl, TDS, DOC, T, and NO_3.

Example time series are shown for two of the wells that sampled 14 of 15 of the indicators of interest, see Figure 7. For the first well, the time series are all of a short duration, since 2012. For the second well, the time series show intervals of significant sampling in the 1970s and 2010s. Indicators that exhibit samples below detection tend to result in noisy time series. Time series with this character are similar among indicators that hover around the detection limit, such as the majority of the PAH indicators.

Conversely, some indicators tend to show atypical values within certain formations (Fig. 8). For example, TDS measures high in well A completed in the Empress formation with occasional detection of low values. Likewise, Cl in well (B) tends to occur at values ranging between 30 and 25 mg/L, with the occasional occurrence of a very low value (5 mg/L). SO_4 in well C tends to be more variable over time with an apparent increasing trend with an occasional sharp decreases. Several factors may be responsible for these uncharacteristic values, including: sample handling (collection, labeling, preservation, storage, and transportation) sample contamination during or after collection, or a typographical error. It is also possible that an extreme event such as a lead to a sudden increase of meteoric water in the wells, shortly before sampling. In examples of Figure 8, the unusual values are lower

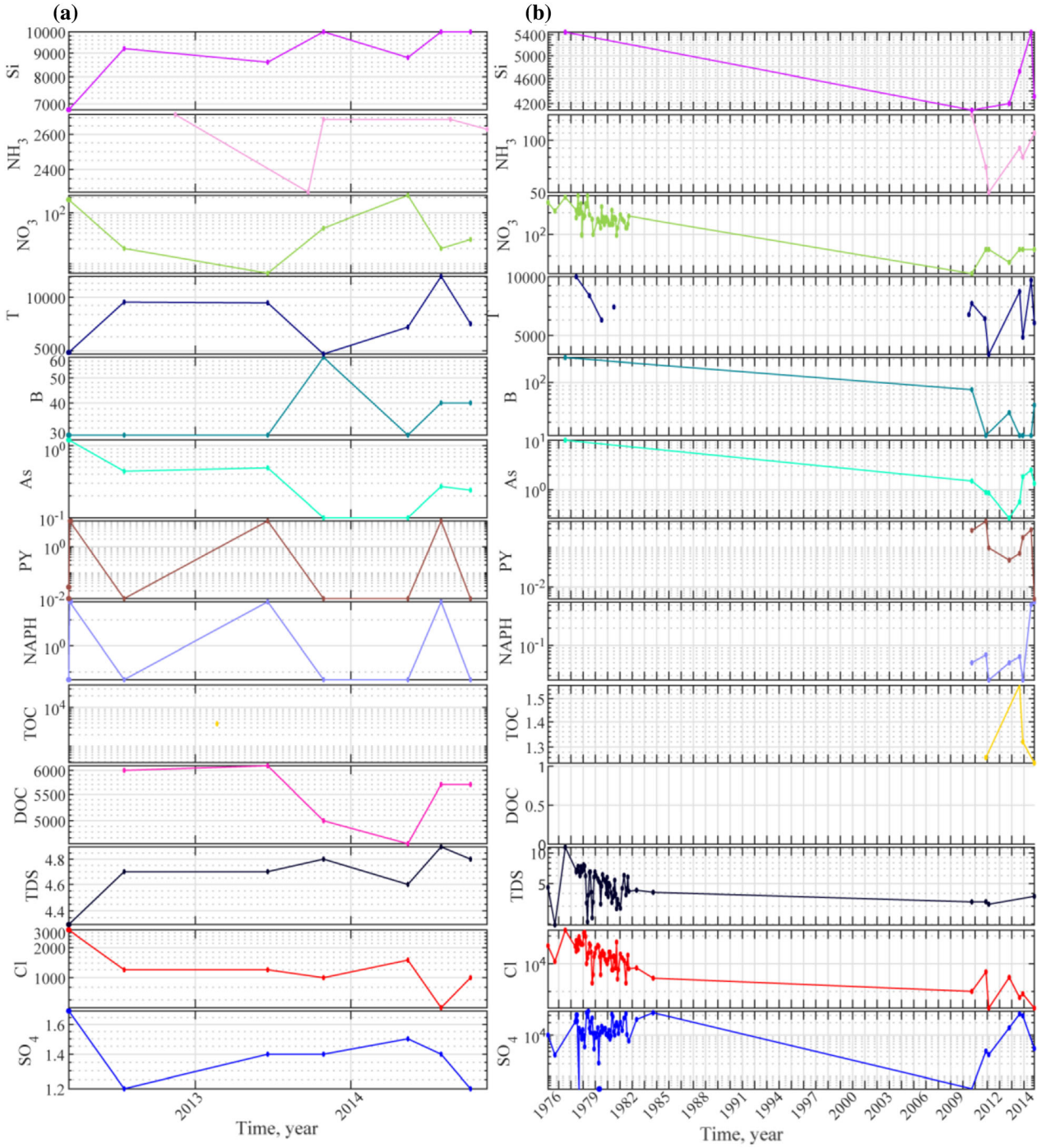

Figure 7. Sampling over time monitoring well "East Christina 76-05-19" that sampled the Sand River formation in the SAOS (left) and monitoring well "GWN-06-60" with two periods of sampling and a period of inactivity for the McMurray formation in the NAOS (right).

suggesting a possible dilution due to an influx of fresher water at the point of sampling. The rapid recovery thereafter is a further indication of a local influx, possibly related to a storm event which is subsequently flushed out of the well during subsequent sampling. Regardless of the cause, such values

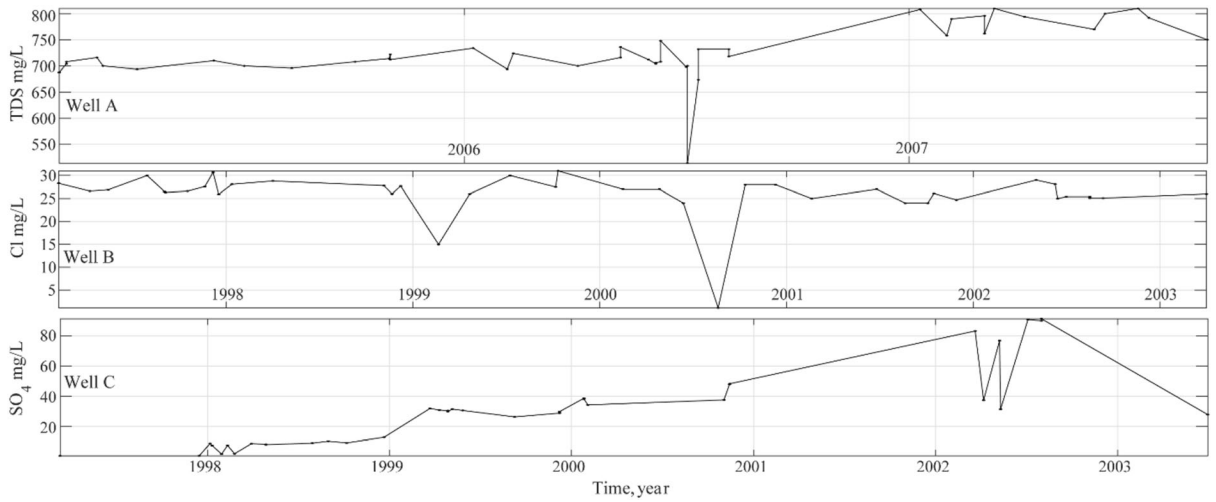

Figure 8. TDS time series from Well A in the Empress formation showing an atypical sample (top); Cl time series from Well B in the Empress formation with two atypical samples (middle), and SO_4 from Well C in the Empress formation with some odd variation between 2002 and 2003 (bottom). All wells were in the CLBR region.

are sporadic and can lead to additional noise, especially for shorter time series. This complication requires the processing these time series with robust methods not affected by outliers.

Sampling variance was evaluated by calculating the temporal nugget effect that yields an estimate of the variability in sampling that occurs over very short time intervals. Such variations are a function of the distribution of ions in water at any given time, sampling practices, laboratory equipment, associated water properties, and other factors. For example, a measurement of Na taken on a particular day is likely to be different if taken again the very next day, or later in the same day with a variability that is quantified by a temporal nugget effect. Estimating the nugget effect was done by calculating the squared difference, $(x(t_i) - x(t_{i+1}))^2$, between all samples, $x(t)$ that are adjacent in terms of time, t, and at the same sampling location followed by fitting the result using linear regression. The function fits the squared difference as a function of time, with the y-intercept being equal to two times the nugget effect since the sum of the squared differences is equivalent to the variogram.

To stabilize the process, the normal score transform was applied to each indicator, and hence, the nugget effect is representative of a standardized value of the original units of each variable (see Table 1). The alternative approach that would estimate the standardized nugget effect from the original units of each variable was not considered since it would be sensitive to rogue values that were previously identified as well as other potential outliers. The normal score transform suppresses the impact that such values would have on the results, rather than having to explicitly filter them from the database.

Table 1. Nugget Effects of the Normal Score Variables of Interest for all Formations Combined

Variable	(Nugget effect)$^{1/2}$ of normal scores	Original units
PY	0.000	µg/l
NAPH	0.113	µg/l
Na	0.069	mg/l
Si	0.152	mg/l
NH3	0.128	mg/l
NO3	0.150	mg/l
T	0.251	°C
NA	0.142	mg/l
TOC	0.207	mg/l
DOC	0.198	mg/l
B	0.131	mg/l
As	0.240	mg/l
TDS	0.071	mg/l
SO_4	0.104	mg/l
Cl	0.110	mg/l

Estimates of the nugget effect in the units of an indicator are representative of the change that could occur over the time span of sampling events, rather than the change that could occur over an infinitesimal change in time. To convert the nugget effect

into the units of the variables prior to a normal score transform that are listed in Table 1 requires knowledge of the mean since the distributions are non-symmetric and generally skewed resulting in a proportional effect; therefore, the conversion in expected value cannot be done to be representative of an entire formation. For the bootstrap approaches that follow, the nugget effects were assumed to be homoscedastic temporally and spatially allowing a dependence on the local mean to be obtained through random sampling and transformation with the global distribution of each indicator, where local refers to any time or spatial location.

Bootstrap-t for Time Anomalies

For water quality monitoring, there is an interest in detecting if an indicator changes in time (i.e., time anomalies). The existence of a time anomaly of statistical significance for a variable measured from a specific hydrostratigraphic unit could indicate a natural change in aquifer conditions that occurred over time due to the dissolution of existing minerals, or other ongoing in situ chemical processes. A time anomaly could also indicate contamination by industry, including contamination from disposal wells, or chemicals leaching from tailings ponds, or by chemical and/or thermal energy transfer from steam injection where cyclic steam or steam-assisted gravity drainage has been implemented during in situ mining operations. Groundwater quality time anomalies could also be induced if surface activities change the quantity or quality of recharge, or if changes in groundwater flow paths result in different degrees of mixing between hydrostratigraphic units with contrasting water chemistry.

Due to the limited length of time series for the indicators of interest, the bootstrap-t technique is determined to be a viable approach for detection of a difference at different times (DiCiccio and Efron 1996). For two intervals of time, such as from 1958 to 1989 and from 1990 to 2015, the bootstrap-t approach is used to conduct a two tailed t-test for different mean values in the two intervals of time. For the sample sizes available, techniques for stabilizing the confidence intervals to obtain better behaved test statistics are utilized (Polansky 2000). Two approaches are used for the water chemistry database and included spatial data aggregation and the use of the smoothed bootstrap (Silverman and Young 1987; DiCiccio and Efron 1996). Because a large number of wells have very limited samples, often less than four, consecutive data from hydrostratigraphic units that are in proximity and that sample the same indicator are considered simultaneously to increase the sample count. Similar to a temporal nugget effect, spatial nugget effect may cause significant changes in water quality within short distances. Therefore, the maximum distance considered for an observation to be included in the series was 500 m, which is approximately half the average well spacing for all wells in the network.

The smoothed bootstrap assumes that each measurement follows a continuous distribution rather than a discrete fixed value. Distributions were generated using the global distribution for each indicator coupled with its corresponding temporal nugget effect (Table 1). To define a distribution for a given measurement, x, with probability, $F(x)$, the estimated nugget effect is used to define a probability interval centered on $F(x)$ from which samples are drawn that take on values from inverting the global cumulative distribution function (CDF). Dissolved chloride is used as an example, for which the global CDF in the units of Cl samples is shown in Figure 9. The standardized nugget effect was estimated as $0.11^2 = 0.012$ for this variable. For a chloride sample that has a value of roughly 10.3 mg/L, the cumulative probability is 0.5. The probability interval that defines the distribution for this sample is given by $0.5 \pm 0.012/2$, or 0.494 to 0.506. The portion of the global CDF that covers this probability range is used to characterize the distribution

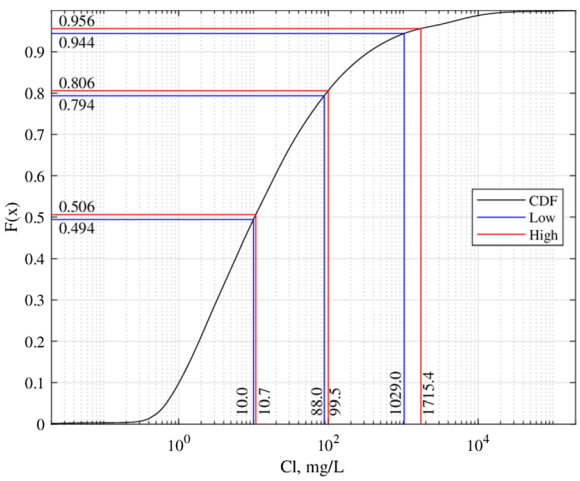

Figure 9. Global CDF of dissolved chloride.

for chloride samples with a value of 10.3 mg/L, which ranges from 10 mg/L to 10.7 mg/L. Similar examples are shown for a 93.5 mg/L sample and a 1290.4 mg/L sample. Since the chloride distribution is positively skewed, samples with higher values will naturally result in wider distributions as shown.

The bootstrap-t approach is a resampling technique that estimates a statistic from a distribution by resampling from the available data, or from a distribution characterized by the data in the smoothed version. The t-distribution, which is usually assumed known under the assumption of a normal distribution, is assumed unknown for the bootstrap-t approach. Rather, the bootstrap-t approach is used to simultaneously build two distributions: one for the statistic to be tested and another for the t-distribution to obtain estimates of confidence intervals. It is also possible to use the percentile-bootstrap method, whereby a confidence interval is obtained directly from the distribution of the statistic; however, with small sample sizes this can be problematic. It should be considered as a viable alternative for future water monitoring network analysis, along with other approaches, to detect time anomalies since different methods may reveal alternate results worth exploring.

The bootstrap-t approach, applied using a significance level of 0.05, proceeds as follows for comparing the mean of an indicator in two intervals of time, given that the null hypothesis is that the mean values are the same:

1. Extract all data for the indicator of interest, sorted by well name and time. Then, for each well:
2. Collect the time series and combine it with nearby wells up to a maximum distance of 500 m (other distances could be used, but it should be kept small).
3. Bootstrap the distribution of the difference in mean values between two time intervals from the time series. The smoothed bootstrap is used. The variance of the mean is also estimated.
4. Compute the average difference, m, and the average variance of the difference, s^2.
5. Compute the t-distribution and extract 0.05 and 0.95 quantiles (t_0 and t_1)
6. Specify confidence intervals as $m - st_1$ and $m - st_0$.
7. Estimate the test statistic, which is the probability that the difference in the mean falls in the 0.05 or 0.95 tails. Values falling in the tails indicate the null hypothesis can be rejected for the specified significance level.

Time anomalies were identified for the indicators of interest and for a minimum time series length of five samples. Even though the smoothed approach makes it possible to draw numerous samples from a distribution that covers a specific sampled value, it is not realistic to compare the mean before and after a point in time from one or two samples in each interval. Based on analysis of sample availability, points in time that will maximize the occurrence of suitable time series before and after each time are 1990, 2003, and 2008. For 1990, the interval width was set as wide as possible to discern if there was a difference between all samples taken prior to 1990 and all those from 1990 and afterward. For 2003, a 7-year window was used, and for 2008 a 10-year window was used. Time windows were sufficiently long so that short-term seasonal effects were averaged out.

Estimating Stable Values

Establishing values at the current time provides a baseline for comparison with samples that are collected in the future. It also provides a spatial dataset for mapping purposes (see Birks et al. 2019) that can be used to guide planning of future wells for expansion of the water monitoring network. Stable values were derived to remove noise that was observed in the time series and quantified via the temporal nugget effects. Challenges for the computation include: 1—sampling was initiated after industrial activity began; 2—some series show high variability; 3—rogue samples that are not necessarily outliers exist; and 4—temporal sampling is not consistent. It is infeasible to derive a baseline value prior to industrial activity; however, estimating a stable value representative of a time as close as possible to 2014 is practical.

To mitigate these challenges, the smoothed bootstrap approach discussed previously was used to compare means at different times and to establish confidence intervals representative of specified time intervals. Such values are referred to as stable measured values, since the smoothed bootstrap approach yields the results less sensitive to noise and outliers.

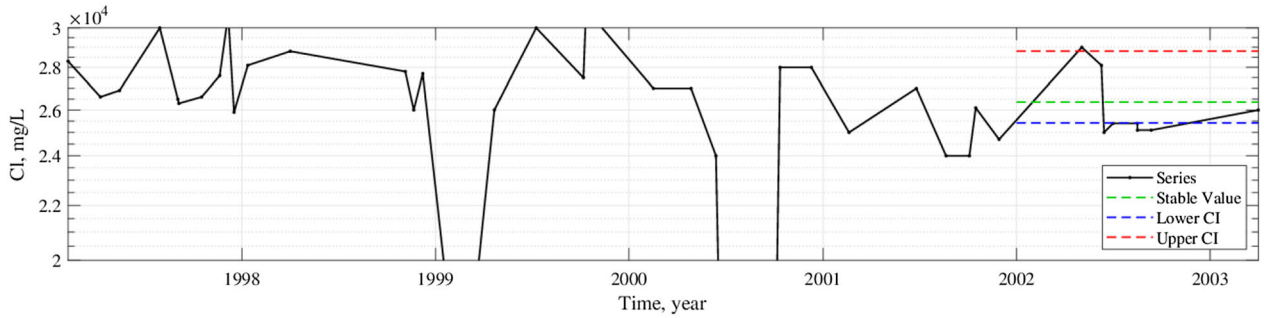

Figure 10. Stable value and confidence interval (CI) from the smoothed bootstrap for the Cl time series of Fig. 7.

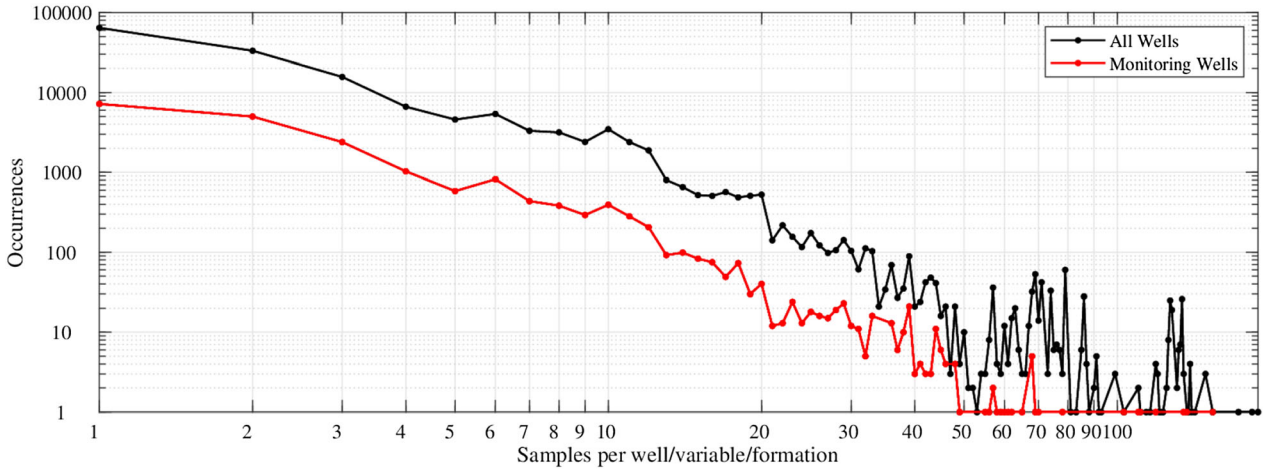

Figure 11. Samples per well/variable/formation for all indicators in the water monitoring database.

For each well and variable, the five most recently available samples were used in the smoothed bootstrap to calculate the stable estimates. In cases where the five samples did not cover at least 1 year of sampling, the set is expanded to the previous year. The resulting stable value and confidence interval for the Cl time series of Fig. is shown in Figure 10. The confidence intervals show skewness that coincides with the range of observations over the most recently available samples and a spread that is consistent with the observed short-term variability of the Cl time series. Future samples that fall outside the confidence interval could be used to flag interesting cases for further assessment.

Not all wells or indicators had five samples available, and not all variables were sampled at equal time increments. A frequency plot of samples per well/indicator/unit is shown in Figure 11, indicating that available samples as a proportion of the database diminishes rapidly. Roughly 40% of the database consists of indicators with a single sample for any given well and formation. Roughly 20% of the database has at least five samples for a given well, indicator, and formation.

Time increments for each indicator per well are summarized in Figure 12. The year of leading samples for pairs and the time increment when the trailing sample was taken are shown. There was a significant amount of samples taken in 2005 to 2007 with time increments less than 1 month, likely from sampling campaigns associated with EIA applications during that period. The most recent 5 years are primarily monitoring wells where most samples are taken between 2 and 6 months apart. Figure 12 reveals high variability of sampling practices. For the most recent five samples, the average interval of time between samples is close to 1 year, while the median is close to half a year.

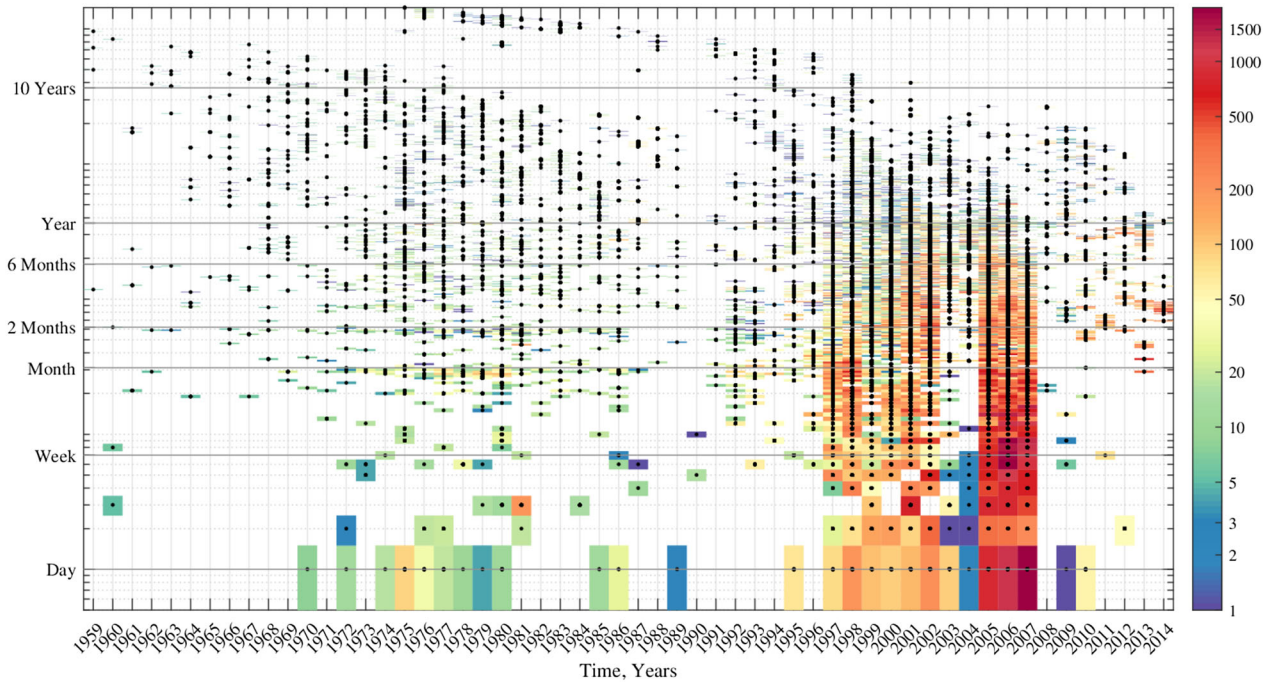

Figure 12. 2D histogram of sample increment frequency by year for all wells and indicators.

As described previously, the smooth bootstrap approach using global distributions sampled over intervals controlled by measurement frequency, and the temporal nugget effect was applied to yield stable values for each well and variable. Afterward, spatial cell declustering (Deutsch and Journel 1998) was applied to account for the irregular spacing of wells for quantifying global distributions of stable values for all variables. Resulting distributions for the indicators of interest are shown in Figure 13.

RESULTS AND DISCUSSION

The results of the water monitoring network analysis were compiled to categorize wells by different properties (type-wells) for the purpose of identifying areas where there is already focused monitoring and areas that may benefit from additional monitoring. To assess the quality of the sampling achieved by the monitoring wells, the average and variance of the cumulative probability of the stable values for each indicator calculated at the monitoring well locations was evaluated. Probabilities were obtained from CDFs of indicators by formation for each monitoring well. Histograms of stable values for four formations are shown in Figure 14 for comparison. Probabilities for all monitoring wells that sample a given indicator in a given formation form the distribution of probabilities that is ideally uniform, which suggests unbiased sampling. The results are meaningful for cases where baseline wells are present along with monitoring wells, which is all cases in this study, otherwise the distribution of probability is guaranteed to be uniform and no information is gained. For a given set of probabilities, a mean of approximately 1/2 and a variance of approximately 1/12 indicates the monitoring wells are centered and well distributed since the probabilities follow a uniform distribution. A high mean or low mean indicates the monitoring wells are primarily sampling the upper or lower tail values observed compared to the global distribution for that indicator and associated formation. When the variance is low, the monitoring wells are sampling similar values. When a high or low mean is accompanied by a low variance, the values are clustered in the tails. As the variance approaches 1/4, the monitoring wells are sampling both the upper and lower tails, since this coincides with the maximum possible variance of a random variable in the

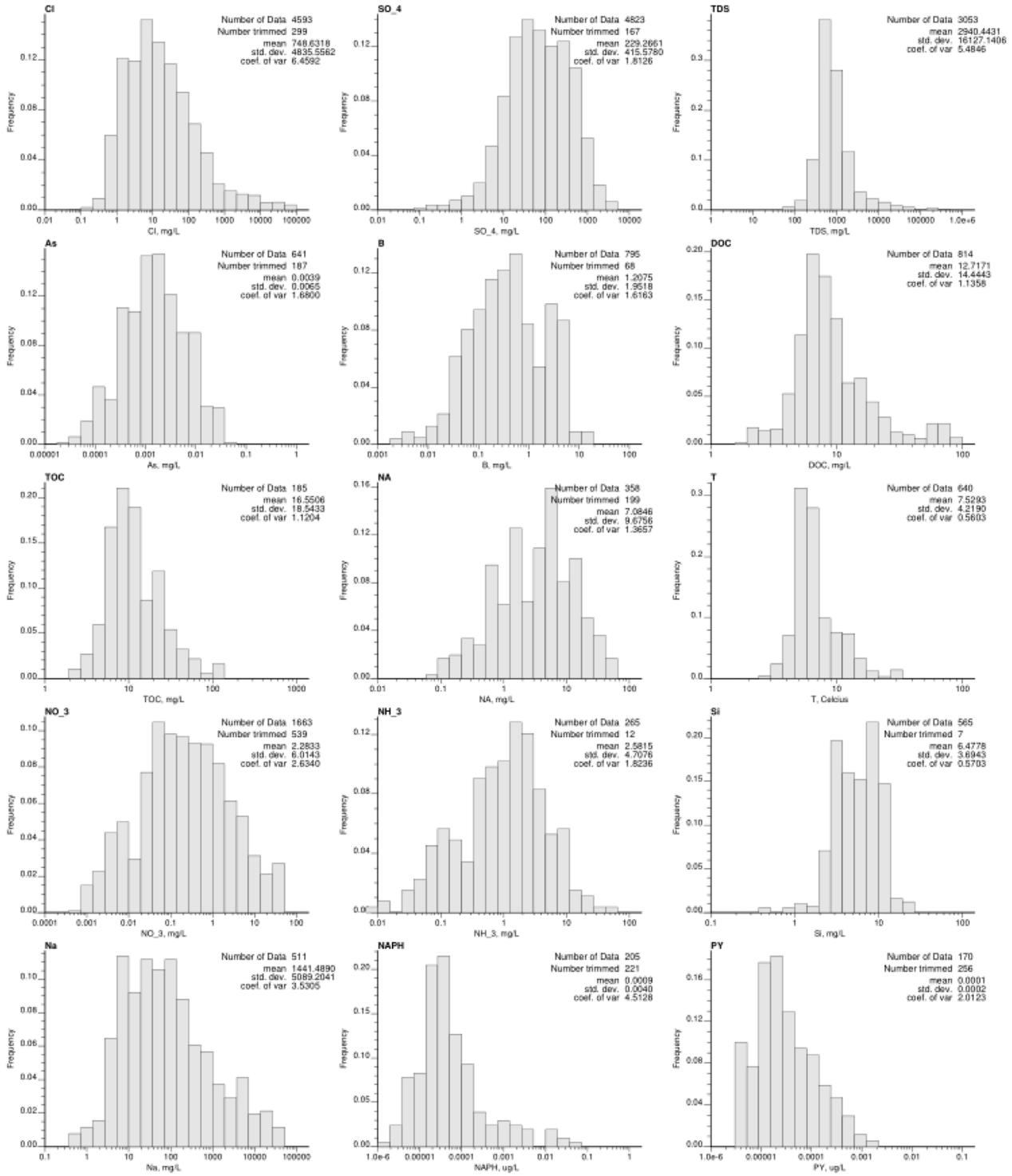

Figure 13. Declustered global distributions of stable values for the indicators of interest.

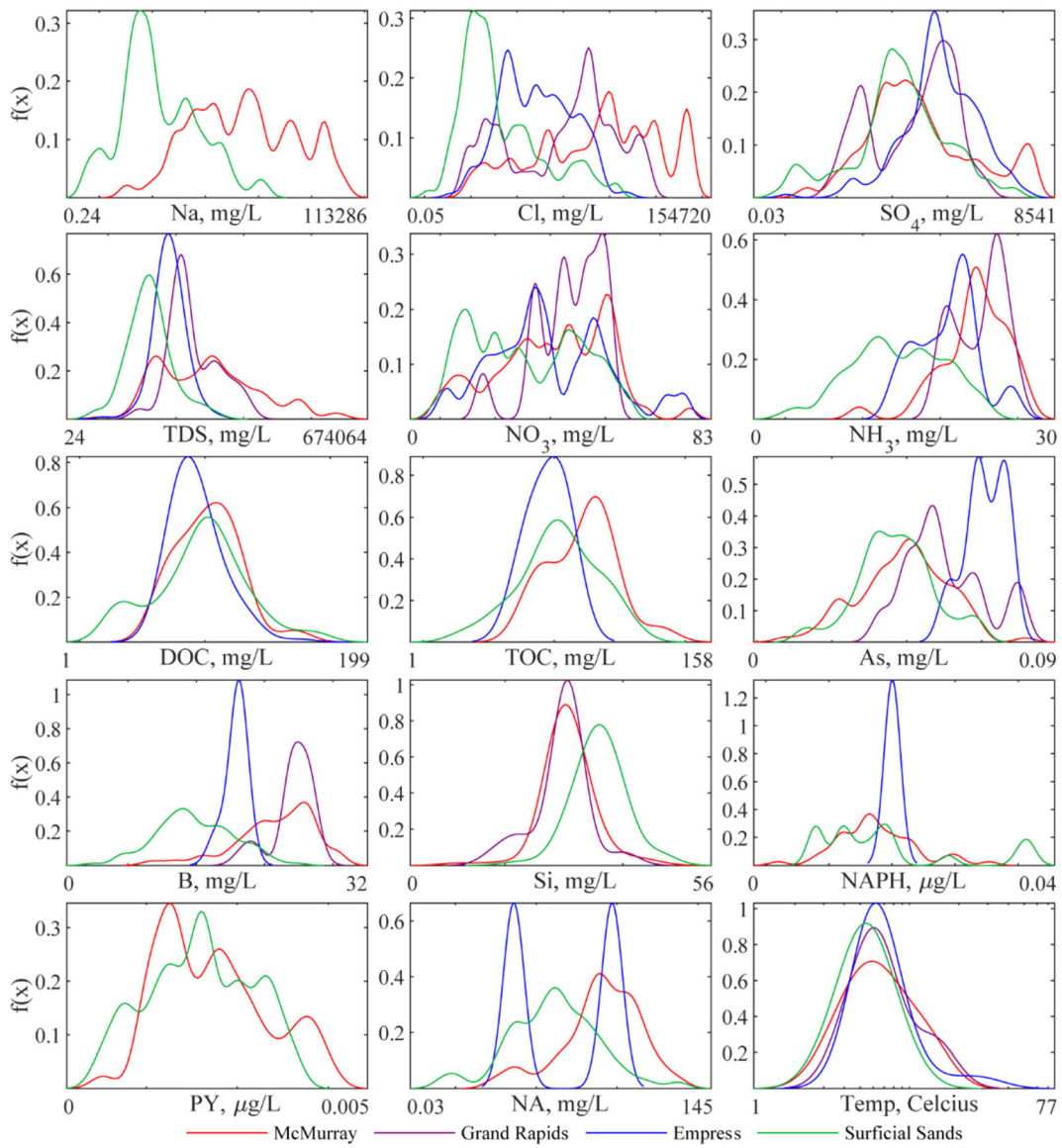

Figure 14. Kernel density estimates of distributions of selected indicators for the McMurray formation, Grand Rapids formation, Empress formation, and surficial sands.

[0, 1] interval that is obtained with an equal proportion of zeros and ones.

The mean and variance of probabilities for each indicator in each formation (indicator–formation pairs) were combined into a sampling score that is defined as the distance that the mean and variance pair is away from perfect uniformity. The distance or score ranges from 0 to 0.5 when the variance is normalized to the [0,1] interval. The 90th percentile is typically associated with a score of 0.3, so any indicator–formation pair that has a score higher than this has sampling bias based on the distribution of the indicator in the formation and on the current monitoring wells that sample that indicator in the formation. An example of several mean versus normalized variance pairs for the indicators is shown in Figure 15, colored by the score variable and also showing contours associated with the 50th, 80th, and 90th percentiles, which coincide with scores of 0.1, 0.2, and 0.3, respectively.

A low score does not necessarily denote good sampling by the monitoring wells since the true

underlying distribution is not known, especially in cases where few baseline wells are available in a given formation along with monitoring wells. However, a high score does indicate that the sampling is likely biased relative to what is known about the true distribution of each indicator from available data. An example of a high score with bias is shown in Figure 16 for the Grand Rapids formation and NH_3 along with a low score without bias in the same formation for Cl.

Resulting scores for the monitoring network for the indicators of interest and formations with monitoring wells present are provided in Figs. 17 and 18. Formations not shown were not being monitored by the AEP well network as of 2014. Scores in the McMurray formation are sufficiently uniform suggesting that the monitoring network may be satisfactory there; however, for the Middle Devonian formation, there is a bias being detected for Cl, SO_4, TDS, As, B, Si, pyrene, and temperature suggesting additional samples should be measured from other wells in that formation. Identification of substantial bias for > 5 indicators, as in the Muriel Lake, Grand Rapids, Beverly Channel, and Empress formation suggests a need to expand the network and sampling for such formations.

Biases may identify additional information about the monitoring wells beyond sampling coverage, especially if monitoring wells are clustered and the concentration of indicators is changing over time. This would place the samples from monitoring wells into one of the tails of the global distribution. Identifying indicators and formations where time anomalies were detected are also provided in Figure 19. Across the 2008 time of interest (Fig. 14), for the McMurray formation, sampling bias is not identified, but 9 of the 15 indicators of interest are showing 1 to 5 time anomalies. This suggests that the magnitudes of the anomalies are within the bounds of the data, thereby providing time to investigate prior to an exceedance of the bounds of the existing data. None of the 2008 anomalies are showing upward changes where there is also a sampling bias. There is an anomaly with a downward change for NA in the surficial sands, so regardless of which tail the bias is concentrating, the decrease in NA is a positive observation, but may be associated with changes in assay methods. For CLBR the surficial sands, there are also increasing anomalies for Cl. Five increasing anomalies for Cl were identified in 2008 for the McMurray formation, four for B, three for TDS, and two for NA, while four decreasing anomalies were found for SO_4. The 2003 anomalies (Fig. 18) also do not show any increasing anomalies coupled with bias, while four anomalies for increasing DOC, and three with increasing NO_3 in Ethel Lake were identified. In the McMurray for-

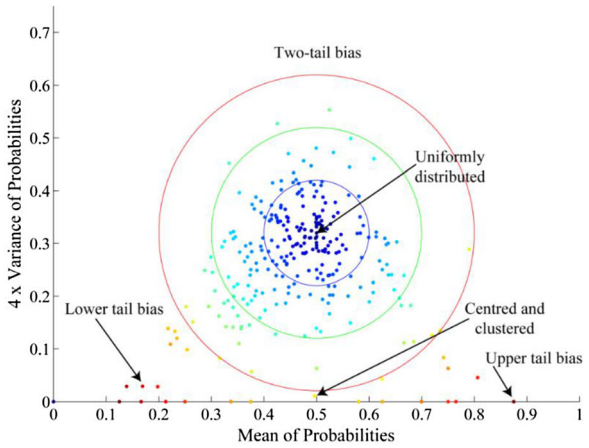

Figure 15. Sampling score plot for testing distributions captured by monitoring wells. Contours are the P50 (blue), P80 (green), and P90 (red) of the score quantity.

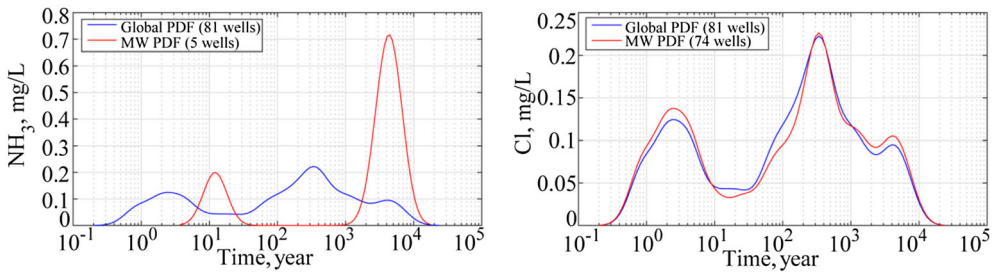

Figure 16. Example of a case with a high sampling score and bias (top) and a low sampling score with no bias (bottom) between the global distribution of stable values from all available data for the Grand Rapids formation and from the MW alone. Distributions were estimated with kernel density estimation for display purposes.

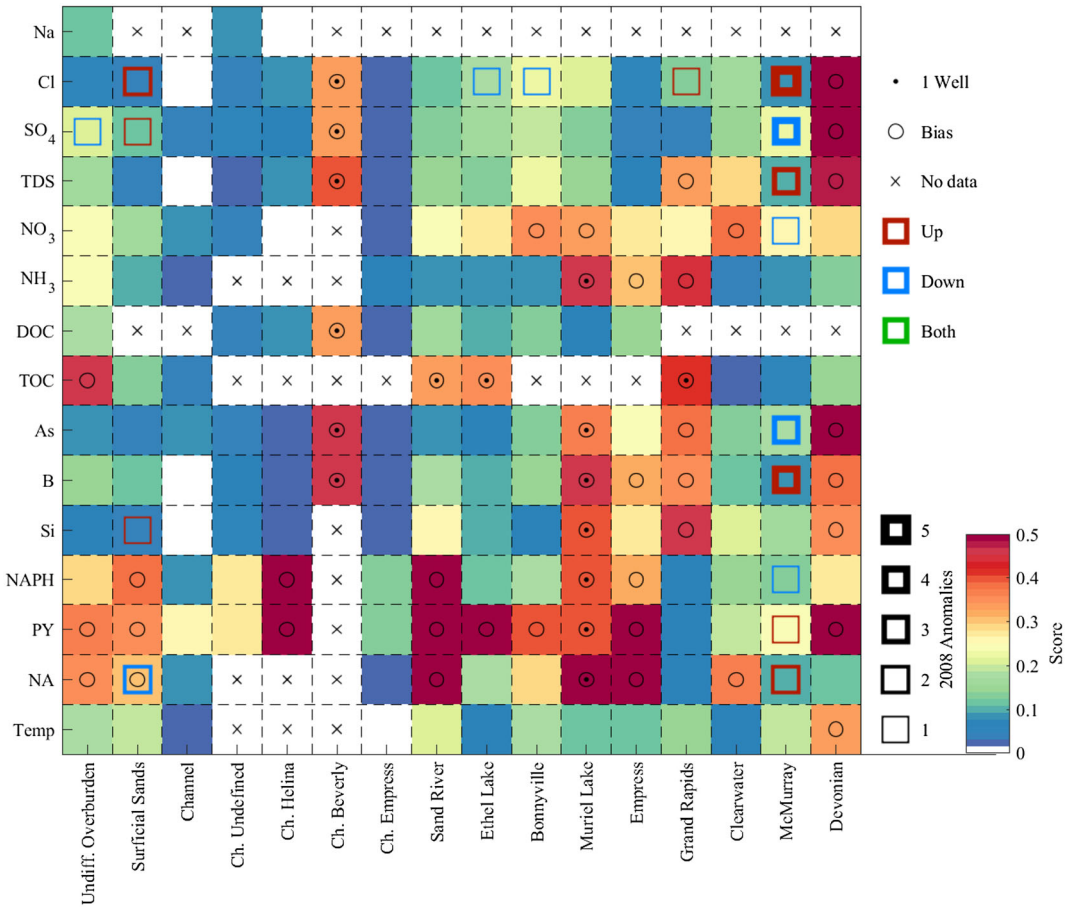

Figure 17. Sampling scores and 2008 time anomaly occurrence for AEP's active monitoring wells by formation and for the indicators of interest.

mation across 2003, similar to 2008, increasing anomalies were found for Cl, TDS, B, and NA, and a decreasing anomaly for SO_4.

Anomalies for 1990 (Fig. 19) show a large number of decreasing anomalies between measurements taken prior to 1990 and after that time. This could be due to a change in the nature of activities occurring in the region. Prior to 1990, a lot of construction activities occurred in the region, often involving forest clearing which exposed the land surface resulting in erosion and change in the redox conditions in the subsurface, resulting in subsequent mobilization of ions. After 1990, most of the construction work was completed resulting a progressive reversal toward initial conditions. Looking back at the time series example in Figure 7, the points prior to 1990 are notably higher than those after 1990. For all anomalies together, trends can be identified such as Cl and TDS in the McMurray Formation that show a reduction from prior to 1990 followed by increases across 2003 and again across 2008.

The results of the analysis were compiled to group monitoring wells into various types. Well types are patterns of results that indicate a measurement of interest, or that the well is situated in a location or formation of interest. Deriving a single type categorization is non-trivial since the wells record many different indicators and intersect one or more formations. The results of interest include the following:

1. Monitoring well sampling *ordination* to explain where the samples from a monitoring well are located in the global distribution for a given indicator and associated formation.

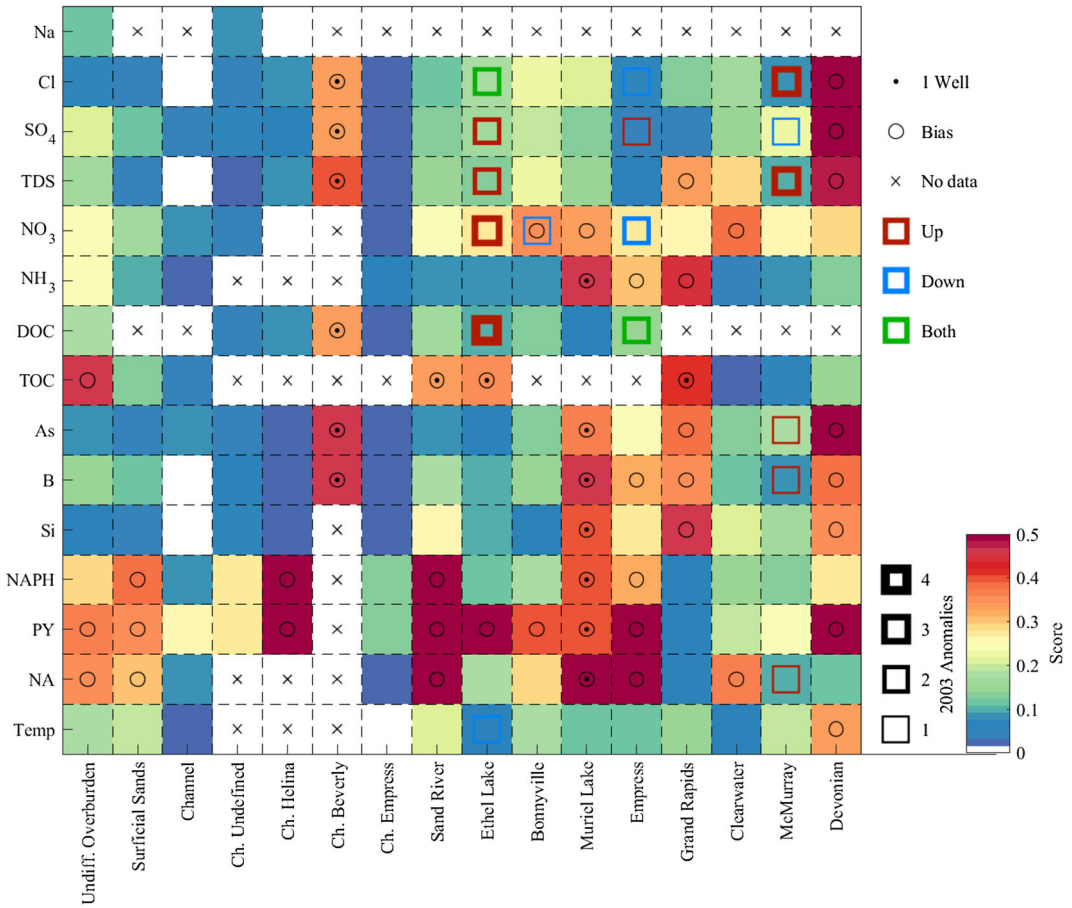

Figure 18. Sampling scores and 2003 time anomaly occurrence for AEP's active monitoring wells by formation and for the indicators of interest.

2. Monitoring well sampling *entropy* across indicators and formations to isolate wells with consistently high or low measurements, suggesting that they are important to the network.
3. *Spatial position* of monitoring wells relative to industry activity to identify the potential for industry driven changes.
4. Presence of recent *time anomalies* to identify if an indicator is changing.

The first component, ordination, is evaluated by computing the quantile for the stable values of each monitoring well relative to the global distributions within each formation and for each indicator and subtracting 0.5 from the result, which yields a deviation from the median. Deviations were also represented by zeros if the measurements were in the lower tail and ones if the measurements were in the upper tail, yielding a binary series. The proportion of zeros and ones were used to compute the Shannon entropy, the second component, given by Eq. 1, where p_0 is the proportion of times a well samples below the median (zeros) and p_1 is the proportion of times above the median (ones). Shannon entropy has a maximum of approximately 0.693 when $p_0 = p_1 = 0.5$. When the entropy is below 0.5, 80% of the samples are measuring either above or below the median, and this value is used to classify wells as having high or low entropy.

$$H = -p_0 \log(p_0) - p_1 \log(p_1) \tag{1}$$

For spatial positioning, the distance from a monitoring well to an oil sand-related project boundary was calculated. Because no information was available about where activity was taking place

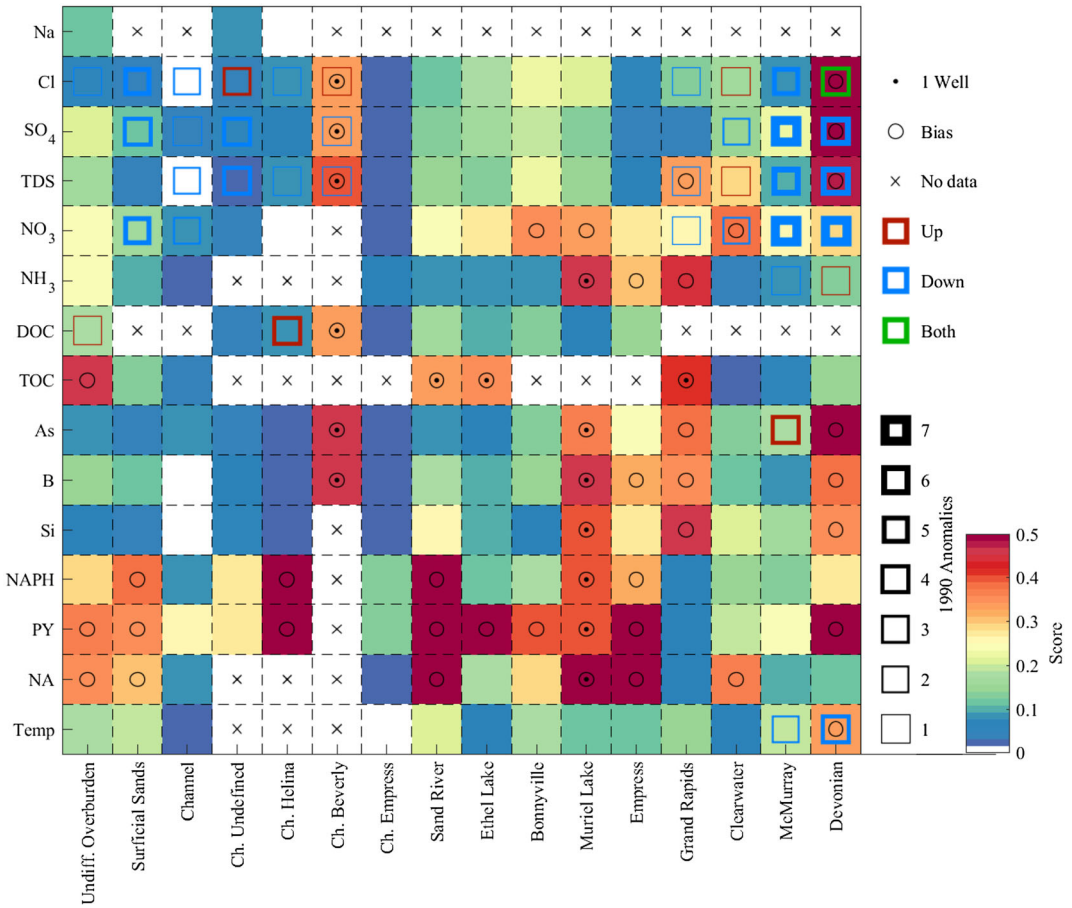

Figure 19. Sampling scores and 1990 time anomaly occurrence for AEP's active monitoring wells by formation and for the indicators of interest.

within project boundaries, no assumptions related to the proximity of a well to some feature within a boundary were made. For example, the distance to the centroid of a boundary has no indicated importance because the actual industry activity may not be at such a location; therefore, any spatial trend observed for an indicator inside project boundaries may be artificial. The signed distance to a boundary was calculated and compared with the other components, deviation from the median (ordination), entropy, and time anomalies across 2008 in Figure 20, where negative distance is inside a boundary and positive is outside. There is a possible trend among the low entropy points as distance increases, which is likely due to regional changes in water quality. Most of the time anomalies are in the NAOS inside the project boundaries and may be associated with well age; a few low entropy points are sampling above the median inside project boundaries. Roughly 2 km outside the project boundaries, there is a cluster of points associated with low entropy wells sampling the lower tails of the global distributions that are well situated for future monitoring.

Well types are provided on a map in Figure 21, indicating that all anomalous wells detected for the 2008 time interval are inside, or in close proximity to project boundaries. The cluster of low entropy wells falling in the upper tail of the stable value distributions are clustered in the Southwest portion of the map, with no apparent explanation for the difference apart from regional differences across a large area. These wells are completed in the Ethel Lake, Sand River, and Empress formations. As mentioned, many low entropy wells fall within close proximity to project boundaries that are consistently sampling in

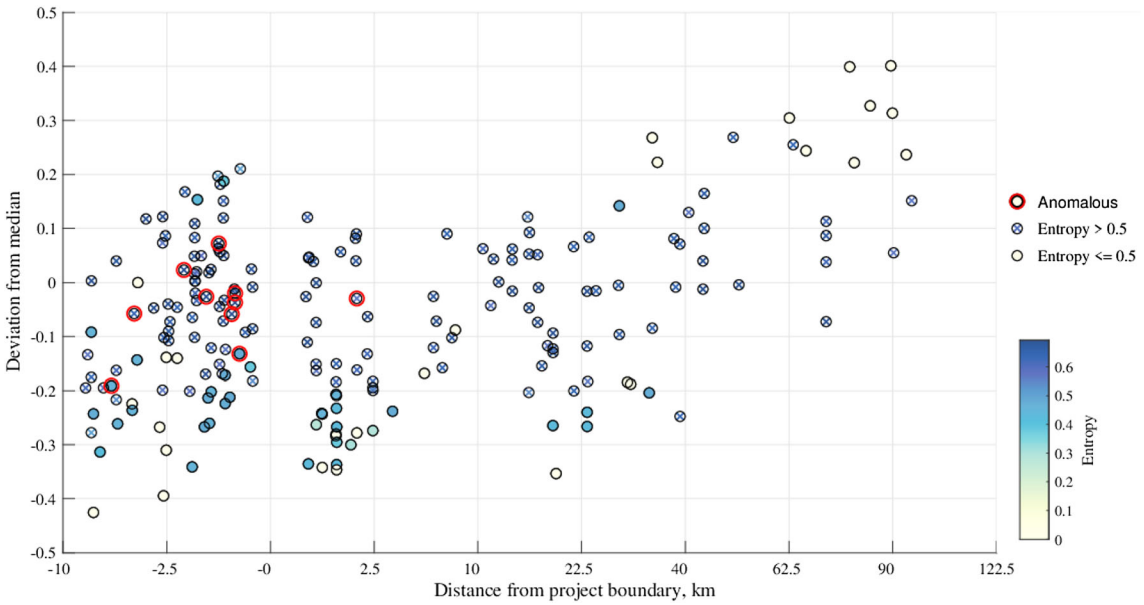

Figure 20. Deviation from the median of stable values for AEP's active monitoring wells related to distance from oil sand project boundaries. Positive distance indicates wells outside the boundaries, while negative distance indicates wells inside boundaries. Time anomalies shown are across 2008.

the lower tails of the distributions. A possible explanation for this is that depressurization due to ground removal from mining, dewatering, and water usage from oil sands related activity is resulting in a hydraulic gradient driving groundwater flow toward the projects. Low entropy wells sampling the lower tail of the distributions are good candidates for continued observations because they will likely reveal statistically significant time anomalies in the future should an increase in the measured indicators occur.

A cluster of time anomalies in water quality occurs in the central portion of the NAOS in the McMurray formation and surficial sands. These anomalies occur near surface mining, adjacent and east of the Athabasca River near locations where natural saline groundwater discharges to the Athabasca River (Birks et al. 2018), and the McMurray and Devonian formations have high solute concentrations. While these changes could be geochemical indicators of changes in the relative proportions of mixing between formations with contrasting chemical compositions in these areas with vertical connectivity, the cause of these observed increases has yet to be determined. Also note that in the CLBR, time anomalies in DOC occur in the Ethel Lake Formation across 2003 in an area of thermal in situ oil sands development near the confluence of multiple channel formations. Moncur et al. (2015) found DOC concentration increases upon heating of Quaternary sediments; however, the DOC increases observed in this dataset could be due to other natural or anthropogenic causes and requires further investigation.

Some recommendations coming out of the research presented in this paper include the following. (1) Continue to collect long-term groundwater quality monitoring data on a regular basis, for a wide range of chemical parameters, in the oil sands regions from both government monitoring wells and other regional data sources to extend time series and run lengths. (2) Baseline wells, in addition to monitoring wells, contribute valuable information to understand regional groundwater quality, better define the global distribution of indicator parameters for each formation, and evaluate whether the monitoring well network is unbiased. (3) Consider expanding groundwater monitoring coverage in under monitored formations such as Beverly Channel, Grand Rapids, and Empress formations and in areas near active leases. (4) Recommend confirming temporal anomalies in water quality and investigating the cause.

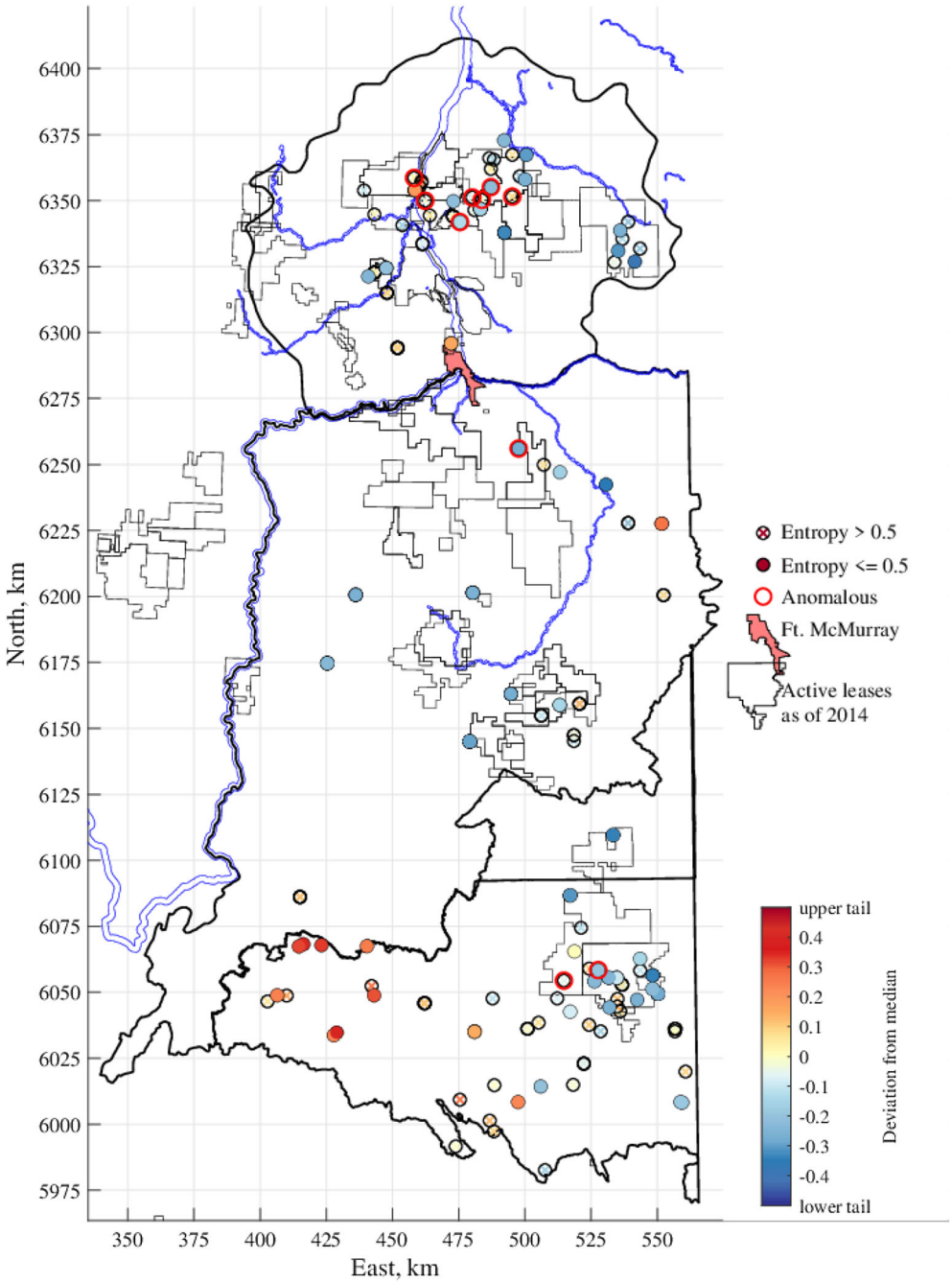

Figure 21. Well types derived from the groundwater monitoring network analysis in the NAOS, SAOS, and CLBR. Time anomalies shown are across 2008. Lease boundaries accessed from: http://osip.alberta.ca/library/Dataset/Details/729, March 2019 accessed from: http://osip.alberta.ca/library/Dataset/Details/729, March 2019.

CONCLUSIONS

Statistical analyses including smooth bootstrap, bootstrap-t, entropy, and deviation from the median were utilized to evaluate the groundwater quality monitoring network implemented in the Athabasca oil sands region of Alberta. Time series of measurements were generally short with various outliers

and noise requiring stochastic sampling strategies and smoothing to yield filtered results for interpretation. Causes for the particular character and quality of the time series are likely economic and political in nature and also due to variation in industry monitoring involvement (e.g., EIA's), ownership, and management of operations. Variation led to the choice of a smoothed bootstrap approach and a bootstrap-t approach to, respectively, derive stable values representative of the sample sites and for detecting changes at specified points in time. The smoothed bootstrap applied to time series helped with smoothing noise, while the bootstrap-t approach was useful in detecting changes in groundwater quality over time from stable values generated by the smoothed bootstrap. Stable values were further used to calculate the Shannon Entropy of time series relative to global distribution of indicators by formation, the results of which were used to isolate monitoring locations consistently displaying extreme values (low or high). Stable values have applications in mapping and as a baseline for comparing the future monitoring results, while type-well analysis that involved entropy and deviation from the median has applications for selecting future monitoring sites and identifying wells of interest.

The most recent time selected to detect if a change was present in the time series was 2008, which showed a cluster of temporal anomalies in the McMurray formation and surficial sands (e.g., increasing Cl) within active oil sands mining leases in the North Athabasca Oil Sands region and relatively close to the Athabasca River. The results of the analysis were combined with a measure of entropy to identify wells of importance, or areas for additional monitoring. An assessment was conducted of the overall deviation of measurements from the median, as a function of distance from active oil sands leases, but no significant trend was observed. Because this considered all variables combined, high entropy of measurements relative to the median were observed in many cases and the probability of a consistent high measurement across all variables was low. Nevertheless, there were numerous cases within 2 km of lease boundaries with low entropy that measured consistently low, and a few cases in the Cold Lake Beaver River region with low entropy measuring consistently higher that may be considered for further investigation. To further assess the coverage of AEP's active monitoring network, sampling scores were assigned and summarized in figures by formation and indicator, with the intention of highlighting formations where a lack of data existed (sampling bias) and where temporal anomalies have taken place. Formations such as Beverly Channel are practically not being monitored in AEP's active network, while others show sampling bias (e.g., Devonian, Grand Rapids, Empress) suggesting a need to expand monitoring in these formations. Others such as the McMurray formation are being actively monitored, but are showing numerous interesting temporal anomalies which require additional analysis. This study demonstrates the use of nonparametric statistical methods to analyze large groundwater quality datasets (> 5000 wells) compiled from multiple sources beyond active governmental monitoring networks (e.g., 161 wells) and confirms that they can inform evaluation of status and trends in regional groundwater quality and monitoring network expansion.

ACKNOWLEDGMENTS

This work was funded under the Oil Sands Monitoring Program, of the Government of Alberta (Alberta Environment and Parks) and Environment and Climate Change Canada, and is a contribution to the Program but does not necessarily reflect the position of the Program. A significant amount of data collection and processing was done by Don Jones of InnoTech Alberta to supplement the work.

REFERENCES

AEP. (2012). *Lower Athabasca region—Groundwater management framework*. Alberta Environment and Parks (AEP).

Akena, A. M., & Christian, L. L. (1981). *Water quality of the Athabasca oil sands area, Volume IV: An interim compilation of non AOSERP water quality data*. Prepared for the Alberta Oil Sands Environmental Research Program by Alberta Environment, AOSERP Report L74.

Andriashek, L. D. (2003). *Quaternary geological setting of the Athabasca oil sands (in situ) area, Northeast Alberta, Canada*. Alberta Geological Survey Earth Sciences Report 2002–03.

Anttila, S., Ketola, M., Vakkilainen, K., & Kairesalo, T. (2012). Assessing temporal representativeness of water quality monitoring data. *Journal of Environmental Monitoring, 14,* 589–595.

Ba, A., & McKenna, S. A. (2015). Water quality monitoring with online change-point detection methods. *Journal of Hydroinformatics, 17*(1), 7–19.

Bachu, S., Underschultz, B.H., Hitchon, B., & Cotterill, D. 1993. *Regional-scale subsurface hydrogeology in Northeastern Alberta*. Alberta Geological Survey, Edmonton, AB. Retrieved

April 1, 2000 from https://ags.aer.ca/document/BUL/BUL_061.pdf.
Birks, J., McClain, C., Manchuk, J., Deutsch, C., Yi, Y., Moncur, M., et al. (2019). *Groundwater water quality monitoring near oil sands development: Regional insights from water management*. Submitted to Geochemical Exploration, April 2019.
Birks, S. J., Moncur, M. C., Gibson, J. J., Yi, Y., Fennell, W. J., & Taylor, E. B. (2018). Origin and hydrogeological setting of saline groundwater discharges to the Athabasca River: Geochemical and isotopic characterization of the hyporheic zone. *Applied Geochemistry, 98,* 172–190.
Breitenberger, S., Efrosinin, D., Hofmann, N., & Auer, W. (2018). Comparison of classic and novel change point detection methods for time series with changes in variance. *Electronic Journal of Applied Statistical Analysis, 11*(1), 208–234.
Buhlmann, P. (2002). Bootstraps for time series. *Statistical Science, 17*(1), 52–72.
Camci, F. (2010). Change point detection in time series data using support vectors. *International Journal of Pattern Recognition and Artificial Intelligence, 24*(1), 73–95.
Deutsch, C. V., & Journel, A. G. (1998). *GSLIB: Geostatistical software library and user's guide*. Oxford: Oxford University Press.
DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3), 189–212.
Efron, B. (1979). Bootstrap methods—Another look at the jackknife. *The Annals of Statistics, 7,* 1–26.
Ghosh, M., & Sen, P. K. (1991). Bayesian pitman closeness. *Communications in Statistics-Theory and Methods, 20,* 3659–3678. https://doi.org/10.1080/03610929108830730.
Government of Alberta. (2017). *Memorandum of understanding, respecting environmental monitoring of oil sands development*. http://environmentalmonitoring.alberta.ca/wp-content/uploads/2018/03/OSM-MOU-December-1-2017.pdf.
Hardle, W., Horowitz, J., & Kreiss, J. (2013). Bootstrap methods for time series. *International Statistical Review, 71*(2), 435–459.
Helsel, D. R., & Hirsch, R. M. (2002) Statistical methods in water resources techniques of water resources investigations, book 4, chapter A3. U.S. Geological Survey.
Iwanyshyn, M., Ryan, M. C., & Chu, A. (2009). Cost-effective approach for continuous major ion and nutrient concentration estimation in a river. *Journal of Environmental Engineering, 135*(4), 218–224.
Javed, M. B., & Siddique, T. (2016). Thermally released arsenic in porewater form sediments in the Cold Lake area of Alberta, Canada. *Environmental Science & Technology, 50,* 2191–2199.
Johnson, R. A., & Bhattacharyya, G. K. (1996). *Statistics: Principles and methods*. New York: Wiley.
Kelly, E. N., Short, J. W., Schindler, D. W., Hodson, P. V., Ma, M., Kwan, A. K., & Fortin, B. L. (2009). Oil sands development contributes polycyclic aromatic compounds to the Athabasca River and its tributaries. *Proceedings of the National Academy of Sciences, 106*(52), 22346–22351.
Kendall, M. G. (1975). *Rank correlation methods*. London: Charles Griffin.
Keum, J., & Kaluarachchi, J. (2015). Calibration and uncertainty analysis using the SPARROW model for dissolved-solids transport in the upper Colorado River basin. *Journal of the American Water Resources Association, 51*(5), 1192–1210.
Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statistics, 27*(1), 386–404.
Lauzon, N., & Lence, B. J. (2010). Artificial intelligence techniques as detection tests for the identification of shifts in hydrometric data. *Journal of Computing in Civil Engineering*. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000042.
Lorden, G. (1971). Procedure for reacting to a change in distribution. *The Annals of Mathematical Statistics, 42,* 1897–1908.
Mann, H. B. (1945). Non-parametric tests against trend. *Econometrica, 13,* 245–259.
Matrix Solutions Inc. (Matrix). (2015a). *2014 regional groundwater monitoring and wells rehabilitation program north Athabasca oil sands area regional groundwater monitoring network*. Report Prepared for Alberta Environment.
Matrix Solutions Inc. (Matrix). (2015b). *2014 program report south Athabasca oil sands area regional groundwater monitoring network*. Report Prepared for Alberta Environment and Sustainable Resource Development.
Moncur, M. C., Birks, S. J., Gibson, J. J., Yi, Y., & Paktunc, D. (2015). Predicting the mobilization of dissolved metals, organics and gas generation from aquifer sediments prior to in-situ operations. In *GeoConvention 2015*, Canadian Society of Petroleum Geologists, Calgary, AB, May 4–8.
Nordgaard, A., & Grimvall, A. (2006). A resampling technique for estimating the power of non-parametric trend tests. *Environmetrics, 17,* 257–267.
Palmer, M. D. (2001). *Water quality modeling*. ISBN: 978-0-8213-4863-5. https://doi.org/10.1596/0-8213-4863-9..
Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*(1), 100–115.
Parker, J.M., & Tingley, K.W. (1980) *History of the Athabasca oil sands region, 1860 to 1960's, Volume I: socio-economic developments*. University of Alberta.
Polansky, A. M. (2000). Stabilizing bootstrap-t confidence intervals for small samples. *Canadian Journal of Statistics*. https://doi.org/10.2307/3315961.
Ranger, M. J., & Gingras, M. K. (2003). *Geology of the Athabasca oil sands—Field guide and overview*. Calgary: Canadian Society of Petroleum Geologists.
Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics, 8*(3), 411–430.
Rodvang, S. J., Mikalson, D. M., & Ryan, M. C. (2004). Changes in ground water quality in an irrigated area of southern Alberta. *Journal of Environmental Quality-Ground Water Quality, 33*(2), 476–487.
Siegmund, D., & Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics, 23*(1), 255–271.
Silverman, B. W., & Young, G. A. (1987). The bootstrap: To smooth or not to smooth? *Biometrika, 74*(3), 469–479.
Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics, 16*(2), 117–186.