User Guide (2st edition, January 2025)

General Application for Groundwater Analytics (GAGA)

Intellectual Property Protection Notice

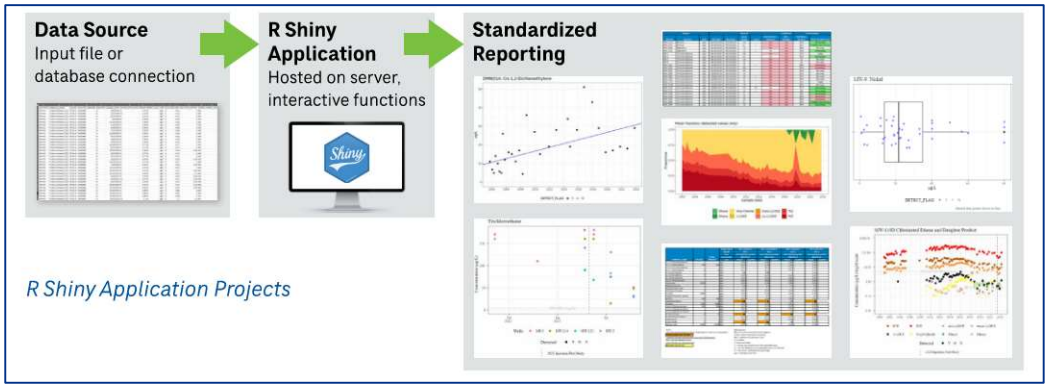
The General Application for Groundwater Analytics (GAGA) tool is for CDM Smith internal use only. ***Please do not share this link with others, including other CDM Smith employees***. You have been provided access to the tool specifically for your project. No part of it may be shared or produced for external viewing without the permission of the Research and Development (R&D) team contacts, the appropriate Business Technology (BT) Data Engineering representative(s), and Office of the General Counsel (OGC). Because the GAGA tool is in development, users need to be tracked for quality assurance until it is ready for broader release.

This application is the intellectual property of CDM Smith, and the use of this application for a client project does not grant the Client any ownership or license in the application. Please communicate with R&D team contacts regarding sharing with other CDM Smith employees or limited client viewing (e.g., slides, screen-sharing). Outputs (charts, screening and statistical tables) can be shared with clients, but it is very important to maintain ownership to the GAGA and that clients cannot access the GAGA. Please work with OGC to ensure we have the appropriate IP protection language.

By accessing this link to GAGA, you agree that you have executed, delivered, are bound and accept the CDM Smith Employee Confidentiality, Invention and Writing Agreement and all CDM Enterprise Intellectual Property and Invention Development policies. For clarification, these terms supplement, and do not modify or replace the terms of the agreements referenced above or any other agreements that you have already entered into with CDM Smith. In the event of a conflict between those agreements and the terms set forth herein, the more restrictive terms shall apply.

Description

The GAGA is an R Shiny web application hosted on an internal CDM Smith server, developed under CDM Smith's R&D program in partnership with CDM Smith Australia. The GAGA is a groundwater data analytics tool that generates automated, accurate, reproducible outputs to support efficient and interactive data analysis and reporting. Appropriate use of this tool can significantly reduce time, budget, and personnel requirements and improve quality of deliverables. Its modules allow the user to upload project-specific data, screen data against numerical criteria, generate a variety of interactive and downloadable time series graphs, and perform statistical analysis in accordance with relevant U.S. Environmental Protection Agency (EPA) guidance and industry practices. Results graphics and tables can be downloaded in the form of consistent, formatted, and organized Microsoft Word, Excel, or png files.



R&D Team Contacts

Emma Ehret, Environmental Engineer – Application Design, Technical Direction, User Experience, Project Implementation – CDM Smith, ehretele@cdmsmith.com

William Lai, Senior Data Scientist – Application Development, Programmer, GitHub Repository Owner – CDM Smith Australia, laiw@cdmsmith.com

Hannah Rolston, Environmental Engineer – User Experience, Project Implementation – CDM Smith, rolstonhm@cdmsmith.com

General Process

1. **Privacy**: Ensure connection to the CDM Smith VPN.
2. **Application Link**: Click/enter link in your browser. Production Version (Released Jan 9, 2025): General Application for Groundwater Analytics (GAGA). This will direct you to the GAGA homepage.
3. **Prepare Data**: Download a template file from the GAGA homepage. Prepare your data to adhere to the template formatting. Save and close the data input file. For more information, see "Data Requirements."
4. **Load Data**: Load the data input file into the application where prompted on the homepage, which will read in your data and supporting organization fields (i.e., well groups, action levels). If GAGA cannot read/load your file, you will see error messages, and you will need to revise your data input file to adhere to the template. For more information, see "Data Requirements."
5. **Use the GAGA**: Navigate throughout the application modules for desired data analysis. There are 3 primary functions: data screening, data visualization, statistical analysis. All outputs are interactive and downloadable, with user options for date range, parameters, and well groups. For more information, see "GAGA Layout and Function."
6. **Post-process**: Check outputs for desired formats, such as print range, notes, headers and footers. Print to PDF and use in reporting.
7. **Give Feedback**: Fill out the feedback form. This is very important for our ability to track this project's value to our firm and R&D program, improve quality and usefulness, and ultimately deploy this tool at a larger scale. Your input is greatly appreciated.

Data Requirements

It is the user's responsibility to ensure that all data loaded into the GAGA meets the project's quality standards. The GAGA reads in all data provided in the input file and does not perform any data cleaning steps or quality checks. Additionally, it is the user's responsibility to ensure that data in the input file adheres to the template format. The following sections provide more information.

Data Assumptions

For GAGA outputs that will provide best quality and value to your project, the following data assumptions and recommendations are provided for the project team's use in preparing the data input file:

- All data to be evaluated in the GAGA will be included in the data input file, which must adhere to the data input file template. The template is available for download from the GAGA homepage and includes all required data fields with notes on required formats. For more information, see "Data Input Template."
- All data will be reportable and will not include any rejected (i.e., R-qualified) data or any data that the project team does not want read in to graphs and analyses. Quality control data will not be included, except for field duplicates, if desired.
- All data will be validated according to the project's data validation requirements. Interpreted validation qualifiers will be included in the data input file. Nondetect data will be qualified with a U-qualifier. Data with qualifiers that do not include "U", are all treated equally in the GAGA. The J-qualifier will be applied to results that are between the method detection limit (MDL) and the reporting limit (RL). The reasoning for reporting values below the reporting limit but greater than the MDL as estimated (J) is to accommodate low detects that are not within the range of the calibration curve. Qualifiers are not adjusted in the GAGA. Only U-qualifiers are used by the GAGA for analyses. Other qualifiers are printed to tables and visible in interactive charts, but not used for analysis.
- All nondetect results will be included. Nondetect data loaded for analysis will be reported consistently to either the reporting limit (RL) or the method detection limit (MDL). Reporting nondetects to the MDL is often the preference for risk assessment, where screening levels are often very close to detection limits. Reporting to the MDL may also be advantageous for statistical analysis. This is because if the J-qualifier is applied when the result is between the MDL and the RL and the U qualifier is applied when the chemical is not detected below the MDL, then it only makes sense to interpret nondetects at the MDL. If there are no J-qualified results and nondetects are reported at the RL, then nondetects must be evaluated at the RLs, which can diminish the power of the analysis to demonstrate adherence to project specific cleanup or screening standards. It is ultimately the responsibility of the project team to decide whether nondetects are reported as RLs or MDLs, as the GAGA does not substitute different values for nondetects for any of the statistical tests or plotting. For more information on nondetects handling in specific statistical tests, see **Attachment 1** other resources (Helsel 2012, EPA 2022).
- All chemicals/parameters have consistent units. If the database has different units for the same parameter over historical data, then these must be changed by the project team to be a single unit for a single parameter in the data input file. All VOCs should be reported in ug/L.
- All data is from the same environmental medium (e.g., groundwater, soil). The statistical analyses employed by the GAGA are generally most applicable to groundwater data.
- If sample depths are relevant to the analysis and results from different vertical depths are to be evaluated separately, then depths should be notated in the sys sample code DATA tab field.
- Field duplicates must be labeled according to the data input template. The GAGA will analyze the greater of the results between the normal sample (N) and field duplicate (FD), unless there is only one detect value, in which case the detected value will be selected regardless of value. If field duplicates are not provided, then the GAGA will use the value included for the unique location and date result.

- Each individual sample must have a unique SYS_SAMPLE_CODE. For example, a sample collected from MW-0X in September 1995 must have a different SYS_SAMPLE_CODE than a sample collected from the same location in April 2003. This is critical for data screening tables and forthcoming radar plots.
- CAS RNs are included for every parameter.

Data Input Template

GAGA users can download a template for the data input file from the Application Information homepage in the application. Data input files that do not adhere to data template format will not load into the GAGA, and the user will not be able to view any other modules.

The data input template has 6 tabs:

- "READ ME" describes each of the fields and their required syntax for each tab of the data input file. The user should not edit this tab and should use it as a key and reference for all other tabs.
- "CAS RN" includes a list of analytes that are included in specific groups for graphing and may have inconsistent CAS RNs assigned across projects. Parameters are plotted on graphs from the data input file based on CAS RN rather than chemical name because spellings and abbreviations often differ between laboratories. Total organic carbon (TOC) and redox parameters often have multiple identifiers and differing CAS RNs. Therefore, it is the user's responsibility to load data for these analytes with the CAS RNs listed in this tab; otherwise, the GAGA will not read them into certain features. If there is no data for a given CAS RN or a mismatched CAS RN, the GAGA will note that there is no data. The user should not edit this tab.
- "DATA" contains all the data that the user wishes the GAGA to evaluate. Each field is listed in the READ ME tab with associated information and requirements. The user is responsible for populating this tab.
- "CRITERIA" contains all the comparison criteria that may apply to chemicals or other analytes. Criteria may be screening levels, cleanup goals, maximum contaminant levels (MCLs), or other values with a unit that matches the data for the given analyte. These values will plot on graphs as a line for visual comparison to the data and are used in the Statistical Analysis module. The user is responsible for populating this tab; however, populating this tab is not required to use GAGA. If comparison criteria do not apply to the project, leave this tab blank.
- "ANNOTATION_LINES" contains important dates for the project. Up to six dates will plot on graphs as a line for visual comparison to the data. The user is responsible for populating this tab. Populating this tab is not required to use GAGA. If the user does not want to show significant dates, leave this tab blank.
- "WELL_GROUPS" contains all SYS_LOC_CODEs from DATA with their associated group, which will be used in the GAGA for categorizing analyses. Dropdown menus will allow the user to select well groups when entered in this tab. The user is responsible for populating this tab; however, populating this tab is not required to use GAGA. If well groups are not relevant to the project, leave this tab blank.

GAGA Layout and Function

The GAGA was designed to perform three general forms of data analysis: data screening, data visualization in time series plots, and statistical testing. This section summarizes the modules and features of the GAGA in order. In most of these modules, the user may select a unique date range and a specific group of well locations for analysis. All downloadable GAGA work products are generated in editable files with some formatting; however, it is the user's responsibility to further format these to match project-specific formatting and editorial requirements (i.e., complete and relevant abbreviations lists, print ranges, headers/footers).

Home Page

Upon clicking the application link, the user is directed to the homepage, which provides general disclaimers, development information, contact information, and version information for the GAGA. From this page, the user can download the data input file template to use to generate their own data input file. When the data input file is prepared according to the template (see "Data Requirements"), the user can browse for their file under "Upload your dataset". Error messages will pop up if the input file does not match the template requirements and will not be read in. If the data input file can be read in, the "Load data" button will appear.

After the user loads the data input file, the following modules will appear in the left-hand-side menu: Dataset Information, Data Screening Tables, Time Series Graphs, Statistical Analysis, Single Dataset View. These modules are discussed below. Thereafter, the homepage will not show the option to load data again unless the web browser is refreshed.

Dataset Information

This module summarizes the data that the user loaded via the data input file and can be used as a high-level check of inputted data. However, the user is ultimately responsible for the quality and content of the data input file (see "Data Requirements").

This module has the following summary features:

- An interactive bar graph shows the total number of data records for each year of the data timeframe. The user can hover over columns for the exact count and year.
- The Criteria section summarizes the parameters (chemical name and CAS RN) with their criteria, as entered in the data input file. The user can view 10, 25, 50, or 100 entries at once by adjusting the dropdown at the top left and search the list at the top right.
- The Annotation Lines section summarizes the events read in from the data input file. These time points will plot on graphs throughout the application as a vertical line and the label will appear in graph legends.
- The Well Groups section lists the wells read in for each well group.
- The Summary of Chemicals by Location allows the user to quickly view which wells have data for which chemicals. In the future, this feature may evolve to list each analyte group from the data.

Data Screening Tables

This module will generate an Excel file with two or three tabs:

- "Data with criteria", a formatted crosstab table with all results for a given timeframe, highlighted if they exceed criteria.
- If there are results that exceed the criteria, the tab "Exceedance summary" contains a formatted table with only the results that exceed the criteria. If there are no results that exceed the criteria, this tab is not generated.
- "Details", a summary of selected data, including the wells, chemicals, and timeframe assessed.

The user selects the date range and well group of interest. It is best to use this module for a limited timeframe, such as a most recent sampling event. Multiple wells from different groups may be selected. Field duplicate results (i.e., results with an "FD" in the SAMPLE_TYPE field of the data input file) will not be considered in this module.

A formatting key is included at the bottom of the Excel screening table; however, the user will need to add project-specific information and abbreviations.

Time Series Graphs

There are four graphing options currently available in the GAGA, and they are summarized in each subsection below. For all graphs, open symbols represent nondetect results, and filled symbols represent detected results. The x-axis is time in years, where the tick mark on the axis represents January 1 of the given year. All graphs are available for both interactive viewing and downloads as formatted and bookmarked Microsoft Word files. Most graphs have multiple color palette options for the data points and lines in the graphs (**Attachment 2**). Some graphs are also available as single image downloads as a .png.

Chlorinated Ethenes and Geochemistry

The default group of parameters plotted in this module include:

- Water Quality Parameters: Total Alkalinity, pH, Chloride
- Total Organic Carbon (TOC) and Redox Parameters: dissolved oxygen, nitrate as nitrogen, nitrite as nitrogen, sulfate, methane, and oxidation-reduction potential (ORP). See "Data Input Template" for a special note on CAS RNs.
- Chlorinated Ethene and Daughter Products: tetrachloroethene (PCE), trichloroethene (TCE), 1,1-dichloroethene, cis-1,2-dichloroethene, trans-1,2-dichloroethene, vinyl chloride, ethene, and ethane. These parameters are plotted as mass concentrations in linear and logarithmic base10 concentration scales and as detected molar concentrations in linear concentration and proportional scales. The user has the option to plot ethene and ethane with the other parameters or separately.

The time events (i.e., events listed under Annotation Lines in the data input file) are shown on the graphs. The user can select multiple wells or an entire well group for the graphs to be generated for a downloaded, bookmarked Microsoft Word file ("Download" button). Additionally, individual wells can be selected for detailed analysis interactively within the application, where the user can hover over each graph and view pop-up information boxes for each plotted data point. Default color formats for data points are fixed, where warmer colors represent parent and intermediary compounds (i.e., PCE, TCE, and cis-DCE), and cooler colors represent daughter products or benign end products (i.e., ethene, ethane, acetylene).

Total Volatile Organic Compounds

Similarly to other modules, the Total Volatile Organic Compounds (TVOCs) module allows the user to select a well group, well subgroup, and date range for analysis, with options to view graphs interactively and download for document delivery. TVOCs includes all volatile organic compounds (VOCs), not just chlorinated ethenes. Graph types include mass and detected molar concentrations with a normal (linear) or log-10 y-axis scale. Additionally, the user can select a specific color palette (**Attachment 2**). The TVOC graphs can be downloaded as a png file (button at bottom left of graph) or as a Microsoft Word file ("Download" button). Molar masses for each VOC were obtained from EPA's Regional Screening Levels (RSLs) - Generic Tables "Chemical Specific Parameters" table last updated May 2024 or PubChem (https://pubchem.ncbi.nlm.nih.gov/ accessed October 2024).

Selected Wells / Single Parameter

This module will generate time series graphs for a selected analyte for locations from the selected date range in two graph formats. The Single Parameter, Single Graph option shows all data from all locations in a single graph, with user options for the data series color palette. The Single Parameter, Faceted Graph option shows each location's dataset as an individual panel of the graph.

Like other modules, the user can view these graphs interactively or download formatted, bookmarked Microsoft Word files. Additionally, the Single Parameter, Single Graph output can be saved as a png file for easy transfer into presentations, reports, messages, or emails. The graphs show a vertical line for the event entered into the data input file 'Annotation Line' fields.

Inorganics Graphs

This module will generate time series graphs for selected inorganic parameters in interactive format by individual well and in downloadable faceted graphs format by well group, for the selected date range. Concentration units graph labels will adjust based the units of the inputted data.

The default group of parameters plotted in this module are pulled in by CAS RN and include the following: Aluminum, Antimony, Arsenic, Barium, Beryllium, Boron, Cadmium, Calcium, Chromium, Cobalt, Copper, Cyanide, Iron, Lead, Magnesium, Manganese, Mercury, Nickel, Potassium, Selenium, Silver, Sodium, Sulfide, Thallium, Tin, Vanadium, Zinc. The units labeled on graphs match the units of the data input file.

Filled symbols are detected results, and hollow circles are nondetect results. When only the year is shown in the x-axis label, this represents the start of the year. The user may select whether total (T) or dissolved (D) fractions be plotted. There are user options for y-axis scale, annotation line, and color palette.

Statistical Analysis

The GAGA Statistical Analysis module incorporates statistical methods that are consistent with relevant environmental industry guidance (EPA 2009, ITRC 2012), EPA ProUCL documentation (2022), and industry practices. However, the GAGA does not include all approaches and considerations that are presented in these or other resource materials. Calculation of statistical parameters for specific tests should not be interpreted as a determination that these tests are the best fit for the project. It is the responsibility of the user to understand the tests to the extent practicable and to apply the test results in a manner consistent with the project's objectives and regulatory framework. Although the GAGA team can be a resource, the site decision-making is ultimately the responsibility of the project team. Please see "Data Requirements" and **Attachment 1** for more information on nondetects handling (e.g.,

substitution with the highest nondetect result of a dataset), statistical tests, assumptions, and references.

Within this module, the Statistical Notes page provides links to all statistical approaches and guidance references. The Application Statistics page is where the user runs the statistical tests. All tests performed by the GAGA are conducted together, generating a single output Excel table for each group of datasets evaluated. The user may select the default subset of analytes or all analytes in the data input file. In the future, more options for different analyte groups will be built into the application.

The user selects a well group, date range, and confidence level (typically 0.95 is standard). The table may take several minutes to generate, after which the Download button will appear. **Table 1** includes a list of statistical parameters shown in the GAGA tables and their definitions.

Single Dataset View

This module integrates the functions of multiple other modules for a single dataset (i.e., single location and parameter), allowing the user to investigate a single parameter and location dataset in more detail on a single page. It will generate a table with all data results, a time series graph with a Theil Sen trendline, a boxplot, and a statistical analysis summary for the selected date range. This page is interactive, and all generated content can be downloaded in a formatted, bookmarked Word file. The graphs can also be downloaded as individual png files (top right button at each graph). Statistical methods are the same as those presented under the Statistical Analysis module (**Attachment 1**). This is a useful module when there are a select few specific datasets that require scrutiny. It is not efficient to use this module for all parameters and locations at a site for a reporting period.

Quality Assurance

As digital solutions evolve at CDM Smith, so do our quality assurance (QA) processes, documentation, and requirements. The BT data technology team is helping us formalize the QA process for tools like GAGA. The QA process for the GAGA development thus far includes the following:

- Verified accuracy of data read-in from the data input file, i.e., correct action levels, location groups, tabulated and plotted results, dates, location information.
- Verified accuracy of the GAGA tables' automatic conditional formatting, based on standards and other parameter exceedances.
- Verified consistency and functionality in user experience (UX) and application products over a 2-year iterative development process between the user experience (UX)/technical designers and the web application developer.
- Incorporation of existing EnvStats R packages into code, with complete documentation.
- Verified comparable outputs to established EPA's ProUCL statistical tests for Mann Kendall tests.
- Statistical methodology review by CDM Smith environmental statistician(s), summarized in **Attachment 1**., for general accordance with industry practices and correct implementation of test procedures.
- Verified comparable or improved (defensible) outputs compared to Rick Chappell's Excel macro statistical workbooks.

A similar application was built for Fort Hall Mine Landfill (Bannock County client), and it is used by various team members. It was developed by Emma Ehret, Tamzen Macbeth (Remediation Practice Lead), and William Lai and is currently overseen by Hannah Rolston. Select IP from this beta version of the tool has been reused and re-verified for the GAGA.

Specifications

Exhibit 1. GAGA Specifications (CONFIDENTIAL)

Application	General Groundwater
Programming Language	Shiny Application scripted in R language
Code Storage	CDM Smith GitHub repository, archived regularly
Application Host Server	Shiny server on Linux on Amazon Web Services, subscription costs covered by CDM Smith's Enterprise R&D Program
User Requirements	Internet connection, CDM Smith VPN (internal facing)
Use Case	General groundwater data screening, visualization, and statistical analysis
Intellectual Property Reuse	Yes, from existing version
Data Requirements	Formatted input file with data quality and data field requirements. Current maximum size of 16 gigabytes; however, this is adjustable.
Database Requirements	None, see Data Requirements
Outputs	Interactive time series graphs, downloadable and formatted time series graph Microsoft Word files and Microsoft Excel statistical summary tables

Project Status and Feedback

We want to be very transparent with the status of this R&D program project and development of the GAGA. Currently, we are testing the tool with a select few projects. Thank you for your participation in this valuable development phase! This guide will be updated concurrently with the GAGA improvements. We are here to answer additional questions you may have after referring to this guide. Thank you for your participation in this valuable development phase!

Your feedback and questions are vital. This is very important for our ability to track this project's value to our firm and R&D program, improve quality and usefulness, and ultimately deploy this tool at a larger scale. Your input is greatly appreciated. Please complete the feedback form at the following link: GAGA Feedback Form.

Version Record

Document Version	Date	Author	Revisions
1	November 2024	Emma Ehret, Hannah Rolston	-
2	December 2025, January 2025	Emma Ehret, Hannah Rolston	Primarily IP protection language; nondetects handling for MK

Table 1: Statistical Definitions

Pull in from Excel

Abbreviation/Expression	Definition
%	percent
µg/L	microgram per liter
BCa bootstrap	bootstrapping method to estimate confidence limits that corrects for bias and skewness in the distribution of bootstrap estimates, used for data sets with all detected values
Bootstrap	method for estimating uncertainty of a statistic where many recreations of the data set are generated by sampling the measured values with replacement, used for data sets with 1 or more nondetect values
CAS_RN	chemical CAS RN
CHEMICAL_NAME	chemical/analyte/parameter
CL	confidence limit
CL conf	confidence level of the confidence limits
CL method	method used to calculate the confidence interval for the dataset
Confidence Level	confidence level of the Mann-Kendall Trend Test
COV	coefficient of variation, calculated as the dataset standard deviation divided by the mean
Dataset end	most recent date in the analyzed dataset
Dataset mean	mean of the records in the dataset
Dataset n	number of records between the Dataset start and Dataset end
Dataset start	earliest date in the analyzed dataset
Direction	Mann-Kendall trend result
EPA	United States Environmental Protection Agency
GSI Toolkit Trend	Mann-Kendall trend result using confidence levels down to 90% to determine a trend (Aziz et al. 2003, Connor et al. 2012)
J	estimated result
Kaplan-Meier	statistical method to estimate the distribution of censored data (i.e., containing nondetect values)
Latest Q	laboratory qualifier for the most recent result (if any)
Latest Result	most recent result
Latest Result > mean	Determination whether the Latest Result exceeds the dataset mean
Latest Result > Standard	Determination whether the Latest Result exceeds the Standard
LCL	lower confidence limit of a confidence interval
LCL > Standard	Determination whether the LCL exceeds the standard
LCL of the mean	lower confidence limit of the dataset mean
Mann Kendall Trend	the output of the Mann-Kendall Trend Test, which is a nonparametric statistical test used to determine whether a time series has a trend
Max detected	maximum result value for the dataset
MCL	maximum contaminant level

Abbreviation/Expression	Definition
mg/L	milligram per liter
Min detected	minimum result value for the dataset
n	number of data results in the dataset
N (background)	number of records for this analyte in the background well (Cell 2 and Cell 4 Monitoring Well only)
NA	not applicable
NC	not calculated
ND	nondetect
ND %	percentage of nondetect results in the dataset
ND (background)	number of nondetect results in the background well (Cell 2 and Cell 4 Monitoring Well only)
NP	nonparametric
ordinary nonparametric bootstrap	method for estimating uncertainty of a statistic where many recreations of the data set are generated by sampling the measured values with replacement, used for data sets with all detected values
Parameter est. method	method used to estimate the mean and standard deviation, which depends on the number of unique detected values in the dataset
ProUCL Method Trend Direction	Mann-Kendall trend result using confidence level of 95% or greater to make a determination of statistically significant (matches ProUCL, EPA 2022)
p-value	probability of the Mann Kendall Trend Test that S would occur without a significant trend
Q	qualifier
S	Kendall's S
sd	standard deviation of the dataset
sd(S)	standard deviation of Kendall's S
Standard	value of the standard (if any)
Standard source	source of the environmental standard or screening level (if any)
TS	Theil-Sen, nonparametric method to estimate dataset trend line
TS Intercept	Theil-Sen intercept
TS Slope	slope of the Theil-Sen trendline in units of concentration unit per day
U	nondetect result
UCL	upper confidence limit of a confidence interval
UCL > Standard	Determination whether the UCL of the mean exceeds the standard
UCL of the mean	upper confidence limit of the dataset mean
UJ	result estimated to be nondetect
Unit	unit of measurement of chemical concentrations
Z	approximated value of Kendall's S, for datasets with n > 10

Attachment 1: Statistical Methods and Assumptions

Statistical Approach and Disclaimer

The GAGA Statistical Analysis module incorporates statistical methods that are consistent with relevant environmental industry guidance (EPA 2009, ITRC 2012), EPA ProUCL documentation (2022), and industry practices. However, the GAGA does not include all approaches and considerations that are presented in these or other resource materials. Calculation of statistical parameters for specific tests should not be interpreted as a determination that these tests are the best fit for the project. It is the responsibility of the user to understand the tests to the extent practicable and to apply the test results in a manner consistent with the project's objectives and regulatory framework. Although the GAGA team can be a resource, the site decision-making is ultimately the responsibility of the project team.

Statistical Methods

The following sections present general descriptions of the statistical tests conducted in GAGA Statistical Analysis module. All statistical tests described are applied to every dataset run through the application, unless the dataset does not meet specific criteria, which are described below for each test. The statistical calculations are performed using the EnvStats R package (Millard 2013), which incorporates EPA's Unified Guidance (2009), and Kendall R package (McLeod 2022), which produces comparable results to ProUCL (EPA 2022). **Table 1** includes a list of statistical parameters shown in the GAGA tables and their definitions.

Comparison of the Latest Result or Dataset Mean to a Standard

Comparison of a value to a standard is the simplest analysis performed on analytical data from all wells sampled. A standard refers to the values entered data input file Criteria tab and is defined by the project team. Standards can include maximum contaminant level (MCL), site-specific cleanup goals, State standards, or risk-based screening levels. Only one value is permitted per analyte; however, if results must be screened against multiple criteria, the user can simply update the data input file with new values and rerun the application and Statistical Analysis module for as many different criteria as apply. The statistical summary output provides the comparison of both the latest result and the dataset mean to the analyte's standard.

Comparison of Confidence Limits to a Standard

Computation of confidence intervals is the recommended statistical strategy for assessment or corrective action monitoring (EPA 2009) because it defines accepted variability in the data and prevents single outliers from driving decision-making. A confidence interval for the mean accounts for variability in sample data and estimates the true arithmetic average of the underlying population that data are sampled from. Confidence intervals are calculated for a data set to allow the comparison of the entire data set, rather than a single result, to a fixed value (i.e., criterion or standard). Confidence intervals may be calculated for different parameters of a population of potential measurements, such as the mean and variance. Additionally, confidence intervals can be obtained via parametric or nonparametric estimation methods. Some parametric methods require as few as two data points to calculate a confidence interval. Using at least eight data points is generally considered best practice (EPA 2009), although data requirements will depend on the variability of the measurements and the objectives of the statistical analysis. Nonparametric methods can be advantageous for sparse environmental data sets, for which it can be difficult to estimate parameters of the underlying probability distribution.

Testing a data set involves calculation of the confidence limit for the arithmetic mean at a determined confidence level, followed by comparison of the confidence limit to the standard. Assessment-monitoring-level criteria, or success during corrective action, is statistically indicated when the entire confidence interval lies to one side of the standard. The lower confidence limit (LCL) and upper confidence limit (UCL) are used as the lower and upper bounds, respectively, of the confidence interval of the data set. The confidence interval, calculated using a particular data set and estimation method, will contain the true value of the population parameter (e.g., the mean) with a probability equal to the confidence level, provided that all statistical assumptions of the parametric or nonparametric estimation method have been met.

The LCL of the parameter (typically an arithmetic mean) is compared to the standard in assessment monitoring. If the LCL exceeds the standard, there is statistically significant evidence that the standard may not be met. If results greater than the LCL during assessment monitoring are observed, additional evaluation is necessary to determine if a sampling, analytical, or statistical error occurred or if the results are due to an alternate contaminant source.

Conversely, the UCL of the data set mean is appropriate when comparing data to the standard in corrective action monitoring. To achieve statistically significant evidence for compliance (and, therefore, successful corrective action), the UCL must fall below the standard.

To calculate confidence limits (i.e., LCLs and UCLs) in the GAGA, the following approach was used:

- Confidence limits were calculated with a 95 percent (%) confidence interval for data sets that contain at least two distinct detected results. If two distinct detected results were not present in a data set, the confidence limits were not calculated, indicated by "NC" in the generated table.
- Bootstrapping was performed for most data sets. For each data set, 2,500 bootstrap samples (i.e., recreations of the original data set, for each of which an equal number of measurement values are generated by sampling with replacement from the original set of measured values) were taken, and confidence limits were calculated based on the variability of the estimated mean across the bootstrap samples. For data sets with only detected values, ordinary nonparametric bootstrapping was performed to estimate the mean, standard deviation, and confidence limits. For data sets with nondetect, or censored, results, the nonparametric Kaplan-Meier estimator for the mean was applied to the bootstrap samples. For data sets with too few unique detected values to compute the confidence limits, the data set mean and standard deviation were calculated without bootstrapping.

Each resulting value is presented in the statistical table output generated in the Statistical Analysis module.

Trend Analysis

Concentration data is analyzed for statistical trends using the Mann–Kendall trend test with a Theil–Sen trend line. The Mann–Kendall test is commonly used to identify temporal trends of COCs in groundwater wells. The Theil–Sen regression line, which estimates a corresponding temporal trend slope, often accompanies the Mann–Kendall test.

The Mann–Kendall test is a nonparametric analysis that compares each data point to later data points in the same data set to develop a summation statistic S based on the comparisons (EPA 2009). The

magnitude of S illustrates the consistency (i.e., lack of variation) of this trend over time, and the sign of S corresponds to the trend's direction. The cumulative probability (p-value) of the test statistic is used to determine whether the trend is statistically significant at the significance level (alpha) chosen. The confidence level, which differs from confidence limits, is the p-value subtracted from one, as a percentage. The coefficient of variation is the dataset standard deviation divided by the mean.

The GAGA calculates the Mann-Kendall test parameters and provides two different trend results – one that matches EPA's ProUCL approach and one that matches the GSI Toolkit approach. The difference in these approaches lies with the determination at the confidence levels. The GAGA calculates Mann-Kendall parameters and trend determination for both approaches for all datasets per the approach below.

- For the ProUCL approach, the alpha is defined to be 0.05. Therefore, a statistically significant trend is present if the confidence level is greater than 95% for increasing and decreasing results, with a direction corresponding to the sign of S. No trend is statistically significant for confidence levels below 95% (EPA 2009, 2022).
- In contrast, the GSI Toolkit approach has more flexibility in interpretation of sitewide data to evaluate plume stability. In this approach, the Mann-Kendall trend determination uses range for alpha to define probably significant trends where the confidence level is between 90% and 95%. Additionally, the COV is used to distinguish between no trend and stable trend results for datasets with confidence levels below 90% for which no statistically significant trend has been identified (Connor et al. 2012). The designation of "no trend" and "no trend – stable" both occur where the presence of a trend is not statistically identified in the given dataset (Connor et al. 2012).

Note that a result of "no trend" or "no trend – stable" does not necessarily prove that there is no trend. Rather, it only means that the test did not detect a trend. Relatively slow temporal trends may not be detectable with a Mann-Kendall test in low-power scenarios where too few data points are used or if the variability of measurements is high (Yue et al. 2002). These results only imply that there is no trend if the test has adequate power.

The Theil–Sen trend line (a nonparametric alternative to linear regression) takes a median slope of pairwise measurements in a data set and applies it to the entire data set using the median concentration and median sample date (EPA 2009). Theil–Sen trend line calculations produce a trend line through the data set, as well as the associated slope. The trend line is shown on the graphs generated in the Single Dataset View module. The GAGA calculates Theil-Sen parameters for all data sets per the approach below.

To statistically analyze concentration trends in the GAGA, the following approach was used:

- Mann–Kendall and Theil–Sen statistical parameters were calculated for datasets with more than 50% detected results, at least 6 results, and no more than 40 results. If these criteria were not met, parameters were not calculated, indicated by "NC" in the generated table.
- Nondetect results are included in both evaluations as the report detection limit value, either the RL or MDL, as determined by the project team (see "Data Requirements"). This is a substitution method for nondetect results. Per EPA's ProUCL approaches, the nondetect results are not

converted to ½ of the detection limit (EPA 2022). For consistency in nondetect results handling, the GAGA has an option to set all nondetect results in a single dataset to the greatest nondetect reported value (e.g., the MDL), where a dataset contains multiple nondetect reporting limits. In this approach, any detected results less than the highest nondetect result is also replaced with the highest nondetect result.

Future Statistical Tests

The primary advantage of the statistical tests currently in the GAGA is conformance with industry convention and regulatory guidance. The primary disadvantage of these tests is a lack of flexibility. Standard contemporary statistical approaches such as mixed models (Stroup 2012) and generalized additive models (Wood 2006) can leverage larger portions of the dataset to yield tighter confidence limits for the mean and estimate temporal trends averaged over subregions of the site. That is, instead of taking the last few years of data and dividing it into dozens or hundreds of individual siloed statistical tests for each compound and well combination, we can use decades of data and make conclusions about processes happening at a spatial scale spanning well beyond the filter pack of a single monitoring well.

Regulator acceptance may be a barrier to relying on these more sophisticated statistical approaches, but there are multiple papers in the scientific literature where mixed models (Chou 2006; Benson et al. 2007; Shoari and Dube 2017) or generalized additive models (Kyte et al. 2023; Mellor and Cey 2015) have been applied to subsurface contamination datasets. Open-source software packages for applying tests based on these statistical approaches are readily available and easy to use. Therefore, future versions of the GAGA may be revised to incorporate some of these more sophisticated approaches.

References

Aziz, J., L. Meng, H. Rifai, C. Newell, and J. Gonzales. 2003. "MAROS: A Decision Support System for Optimizing Monitoring Plans," *Ground Water* 41, no. 3 (May–June): 355–67. [https://doi.org/10.1111/j.1745-6584.2003.tb02605.x]

Benson, Victoria S., John A. VanLeeuwen, Henrik Stryhn, and George H. Somers. 2007. "Temporal Analysis of Groundwater Nitrate Concentrations from Wells in Prince Edward Island, Canada: Application of a Linear Mixed Effects Model." *Hydrogeology Journal* 15 (5): 1009–19. https://doi.org/10.1007/s10040-006-0153-x.

Chou, Charissa J. 2006. "Assessing Spatial, Temporal, and Analytical Variation of Groundwater Chemistry in a Large Nuclear Complex, USA." *Environmental Monitoring and Assessment* 119 (1–3): 571–98. https://doi.org/10.1007/s10661-005-9044-1.

Connor, J., S. Farhat, and M. Vanderford. 2012. *Software User's Manual GSI Mann-Kendall Toolkit for Constituent Trend Analysis*. Version 1. https://www.gsienv.com/gsi-technical-guidance/?resource_search=mann%20kendall

Gibbons, R. D., D. K. Bhaumik, and S. Aryal. 2009. *Statistical Methods for Groundwater Monitoring*. Wiley.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Pacific Northwest Laboratory.

Helsel, D. R. 2012. *Statistics for Censored Environmental Data Using Minitab® and R*. Colorado: Wiley Publishing.

ITRC. 2013. *Groundwater Statistics and Monitoring Compliance, Statistical Tools for the Project Life Cycle*. GSMC-1. Washington, D.C.: Interstate Technology & Regulatory Council, Groundwater Statistics and Monitoring Compliance Team. https://projects.itrcweb.org/gsmc-1/.

Kyte, Emily, Edwin Cey, Leila Hrapovic, and Xiying Hao. 2023. "Nitrate in Shallow Groundwater after More than Four Decades of Manure Application." Journal of Contaminant Hydrology 256 (May):104200. https://doi.org/10.1016/j.jconhyd.2023.104200.

McLeod, A.I. 2022. *Kendall: Kendall Rank Correlation and Mann-Kendall Trend Test.* R package version 2.2.1, https://CRAN.R-project.org/package=Kendall.

Mellor, Andrea F.P., and Edwin E. Cey. 2015. "Using Generalized Additive Mixed Models to Assess Spatial, Temporal, and Hydrologic Controls on Bacteria and Nitrate in a Vulnerable Agricultural Aquifer." Journal of Contaminant Hydrology 182 (November):104–16. https://doi.org/10.1016/j.jconhyd.2015.08.010.

Millard, S.P. 2013. *EnvStats: An R Package for Environmental Statistics*. Springer, New York. ISBN 978-1-4614-8455-4, https://www.springer.com.

Shoari, Niloofar, and Jean-Sébastien Dubé. 2017. "Application of Mixed Effects Models for Characterizing Contaminated Sites." *Chemosphere* 166 (January):380–88. https://doi.org/10.1016/j.chemosphere.2016.09.087.

Stroup, Walter W. 2012. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press.

U.S. Environmental Protection Agency (EPA). *ProUCL: Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations*. Version 5.2. https://www.epa.gov/land-research/proucl-software, 2022.

EPA. 2009. *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Unified Guidance*. EPA 530/R-09-007.

Wood, Simon. 2006. Generalized Additive Models: An Introduction with R. CRC Press.

Yue, Sheng, Paul Pilon, and George Cavadias. 2002. "Power of the Mann–Kendall and Spearman's Rho Tests for Detecting Monotonic Trends in Hydrological Series." *Journal of Hydrology* 259 (1): 254–71. https://doi.org/10.1016/S0022-1694(01)00594-7.

Attachment 2: R Shiny Color Palettes

All are colorblind-conscious options.

Options with limit of 8 distinct colors/datasets (A New palette() for R - The R Blog (r-project.org):

- R4 is the new default palette (same as "default", starting from R version 4.0.0). Dataset/distinct colors limit: 8.
- Classic Tableau: Default palettes (by Maureen Stone & Cristy Miller) from the popular Tableau visualization software. Dataset/distinct colors limit: 8.

Options without a color limit (HCL-Based Color Palettes in grDevices - The R Blog):

- Set 3, Viridis, Inferno, Mako, Plasma, Rocket (Source: HCL-Based Color Palettes in grDevices - The R Blog)