UCD CSN Technical Information #801C

CSN Data Validation

Chemical Speciation Network
Air Quality Research Center
University of California, Davis

November 30, 2018
Version 1.1

Prepared By: _____ Date: 11/28/2018

Reviewed By: _____ Date: 11/28/2018

Approved By: _____ Date: 11/28/18

UCDAVIS
AIR QUALITY RESEARCH CENTER

DOCUMENT HISTORY

Date Modified	Initials	Section/s Modified	Brief Description of Modifications
11/30/18	NJS	1,2,3,7,8,9,10,11	Rewording for clarity and updating name changes.

Table of Contents

List of Figures

1. PURPOSE AND APPLICABILITY

The subject of this technical information document (TI) is validating the analytical data from the Chemical Speciation Network (CSN). Data from the network are reviewed and validated using a variety of tools. Qualifier flags and/or null codes are applied as appropriate.

2. SUMMARY OF THE METHOD

The UCD analyst uses the UCD CSN Data Management website along with custom software in the R language to perform Level 1 validation. The primary tools for review are summary data tables and comparison figures.

3. DEFINITIONS

- **AQS:** EPA's Air Quality System database.
- **Chemical Speciation Network (CSN):** EPA's $PM_{2.5}$ sampling network, with sites located principally in urban areas.
- **crocker:** A custom software package in the R language that contains the data processing code used to produce, check, and post the final results.
- **datvalCSN:** A custom software package in the R language that contains the data validation code used to collect, compare, and flag the final results.
- **Data Analysis and Reporting Tool (DART):** A web application for environmental data visualization and validation procedures.
- **CSN database:** A SQL Server database that is the central warehouse of CSN preliminary and final data at UCD.

4. HEALTH AND SAFETY WARNINGS

Not applicable.

5. CAUTIONS

Not applicable.

6. INTERFERENCES

Not applicable.

7. PERSONNEL QUALIFICATIONS, DUTIES, AND TRAINING

The UC Air Quality Research Center (AQRC) Data & Reporting Group staff assigned to this project all have advanced training in database programming and database management.

8. PROCEDURAL STEPS

Data validation is performed using the *crocker* and *datvalCSN* R packages, which are developed and maintained by UCD specifically for data processing, monitoring, and validation of the CSN data. Data validation is performed by the UCD data management team on monthly batches of data (a calendar month of sample start dates). Validation occurs in four steps:

1. Collect necessary measurement data (Section 8.1)
2. Compare results (Section 8.2)
3. Review summary plots (Section 8.3)
4. Review summary data via web application (Section 8.4)

UC Davis is responsible for the first two levels (levels 0 and 1) of the data validation process. CSN data delivered to DART is Level 1 data while data delivered to the AQS database is Level 2 data. The levels are as follows:

Level 0

Data at this level are raw data obtained directly from the data-acquiring instruments. These data can be reduced or reformatted, but are unedited and not reviewed, without any adjustments for known biases or problems that might have been identified during preventative maintenance checks or audits. These data may monitor instrument operations on a frequent basis (e.g., ambient temperature of sampler every fifteen minutes). Average times represent the minimum intervals recorded, and these data may need to be aggregated to obtain averages for the sampling periods (e.g., 24-hour average). Level 0 data have not been edited for instrument downtime, nor have procedural adjustments for baseline shifts, span changes, or known problems been applied. Examples of data at Level 0 validity in CSN are:

- 24-hour averaged pressure, temperature, and flow data recorded during sample change procedures;
- XRF raw spectra;
- Sample date and sampling time before consistency checks.

Level 1A

Data at this level have passed several automatic and manual qualitative reviews for accuracy and completeness. The focus of Level 1A validation is to obtain as complete a dataset as possible. CSN Level 1A data validation consists of:

- Adding records for expected filter samples that were never generated or were not used as intended, for completeness purposes;
- Setting status flags to reflect sampler malfunctions, site or laboratory analyst errors, or power outages;
- Reviewing analyst, sample handling laboratory, and measurement laboratory comments to verify consistency between records and correct any typographical errors;

- Verifying analyst, sample handling laboratory, and measurement laboratory comments for questionable records;
- Identifying, investigating, and flagging data that are beyond reasonable bounds or that are unrepresentative of the variable being measured, including:
    - Sulfur/sulfate ion ratios outside the range 0.66 to 1.8 and potassium/potassium ion ratios outside the range 0.5 to 2,
    - Gravimetric or collocated mass/reconstructed mass ratios outside the range of 0.5 to 2,
    - Anion/cation ratios outside the range 0.86 to 2.82,
    - Z-scores of the OC/EC ratios outside the range -1 to 1;
- Examining daily flow rates based on the results of the *flow.check* function that identifies abnormal flow rates and significant variations over 24-hours;
- Setting qualifier flags when deviations from nominal operational settings have occurred (e.g., temperature or pressure outside instrumental tolerances);
- Examining the field blank analyses for evidence of swaps with sample filters;
- Examining the sample analyses for evidence of swaps with other sample filters on different days.

Level 1B

Data at this level have passed additional quantitative and qualitative reviews for accuracy and internal consistency. Objective collocated measurements and internal consistency tests are applied by the analyst. Discrepancies that cannot be resolved are reported to the measurement laboratories and sample handling laboratory for investigation. Data that deviate from consistency objectives are individually examined for errors. Extreme outliers (e.g., -85 °C temperature) are noted with a qualifier AQS flag. Changes to the data (e.g., swapping dates on consecutive samples) are recorded and documented by providing comments to the state, local, and tribal (SLT) agency reviewers. Level 1B time-series data review is conducted on a site-by-site basis using a combination of *datvalCSN* package functions and Shiny web applications. The historical archive of CSN data is used to place new measurement values in context with previous measurements. CSN level 1B data validation includes:

- Comparing sulfur and sulfate concentrations as well as potassium and potassium ion concentrations;
- Comparing organic carbon to elemental carbon for both blank corrected and uncorrected concentrations;
- Comparing anions to cations in a molar mass balance;
- Comparing reconstructed mass and collocated $PM_{2.5}$ mass;
- Examining individual data points identified by the various checks as potential sample swaps;
- Reviewing values of each parameter at the network level for a given month to look for anomalies;
- Comparing parameter values and concentrations with those at other nearby sites, within a specified radius, and across the network for any given month;
- Comparing the analytical data to expectations based on prior years.

Level 2

This level of data review is applied after the submission of concentration files to DART, when the results from the measurement laboratories are made available to the SLT agencies. At this level, the data are reconciled with local events and expertise. Upon completion, Level 2 validated data are sent back to UCD for a final review. This review is to ensure consistency between user comments and data transfers. UCD does not make changes to DART-reviewed data unless requested by the SLT agency reviewers.

8.1 Collect Measurement Data, Perform Checks and Comparisons

In the previous step, laboratory results for elements, ions and carbon results are processed into mass per volume of air and posted into the UCD CSN database (see UCD TI 801B). Additionally, collocated data should be collected from the AirNowTech API web service via file transfer protocol (FTP). All data can be collected using functions in the *datvalCSN* package.

The analyst begins validation by opening an R environment (such as RStudio) and running the following command to collect records for all measured and derived parameters from the laboratories, as well as to begin validation by running automated checks[1]:

[monthData] <- datvalCSN::csn_validate(Month = [‘MM'], Year = [‘YYYY’])

The command form shown here is appropriate for typical validation procedures; more options can be included by changing defaults. To find out more information on this command, run *?csn_validate* in the R environment.

Within the *csn_validate* command, records are collected from the *analysis.Results* table that are pertinent to the month and year specified. Data from the AirNowTech API web service is collected when the function parameter *with.mass* is set to TRUE, whereby a function within the *csn_validate* command extracts the desired data from the posted values on the AirNowTech server. All of the results will be returned in memory to the variable *monthData*. The qualifier and null flags assigned from various checks performed within this function can be posted to the UCD CSN database by specifying *write.flags = TRUE* when executing the *csn_validate* function.

8.2 Data Validation Checks: Automated and Analyst Performed Checks

The validation process begins with automated checks. Each check compares resulting values against pre-defined limits as well as comparable parameters. All automated checks are performed as part of the *csn_validate* command. The resulting data tables from these checks are formatted for further graphical interpretation. The individual checks are described below. As each check is a function within the *datvalCSN* R package, it is possible to execute them outside of the *csn_validate* command, if required. In the following descriptions of the checks, the function name is documented; further details

[1] Text in [brackets] indicates values that can be changed by the user. Other values should be typed as written.

can be obtained for each function from the help documents in the R environment by entering a question mark before the function name in the R console.

Operational parameters such as sample flow rate and operating temperature should have been checked both at the sampling site as well as in the sample handling laboratory. However, there are two checks included within the *csn_validate* function for flow and ambient data (function names: *Flow.check* and *Amb.check*, respectively) to compare reported values with instrumental limits to determine if additional flagging is needed. Both functions use the *filter.Filters* table from the database.

The *Flow.check* function generates a dataframe within the R environment containing the records that have aberrant flow values but have not been invalidated. The null flag "AH" will be applied to all of these filters, which represents "Sample Flow Rate or CV out of Limits" in the AQS database. The null flag "SV" will be applied when the sample volume is out of limits. The ambient check will not invalidate samples based on ambient temperature, ambient pressure, or transport temperature, but instead append an appropriate flag to appropriate parameters notifying end users that these operational parameters were outside the specifications only if the appropriate flag has not already been applied. Specifically, if the ambient pressure value is outside of the 600-810 mmHg range then the "QP – Pressure Sensor Questionable" qualifier will be applied to the ambient pressure parameter only. Further, if the ambient temperature is outside of the -20 to 45 °C range for the URG or -30 to 50 °C range for the SASS/SuperSASS then the "QT – Temperature Sensor Questionable" qualifier will be applied to the ambient temperature parameter only. If the transport temperature is greater than 4 °C then the "TT – Transport Temperature is Out of Specs." qualifier flag will be applied to all parameters for the given filter. As part of the checks performed regarding the correct application of the qualifier flags, the "X - Filter Temperature Difference or Average out of Spec." qualifier will be reduced to just being applied to analytical species parameters for a given filter.

Analytical data are checked using various functions that are automatically run when the *csn_validate* command is executed. The three main checks performed, which exclusively use the data collected from the database, collected during the first step of the *csn_validate* command are for ions (function name: *ions.check*), carbon (function name: *carbon.check*), and a cross-analysis comparison, hereafter referred to as the XRF-IC check (function name: *xrfic.check*). The ions check compares the anions and cations data to aid identification of outlier ions measurements. The carbon check compares organic carbon and elemental carbon to aid identification of outlier carbon measurements. The XRF-IC check compares sulfur and potassium elemental concentrations from the PTFE filter to sulfate and potassium ion concentrations from the nylon filter. The ratio of each corresponding element/ion pair is compared to aid identification of questionable samples. For each of the checks, a summary of the results is produced with outliers indicated. Sample records are not invalidated during this process. A qualifier flag may be automatically applied if comparison results lie outside the predefined limits. If flagged, a comment is added to the record to indicate which check produced the outlier flag. An example of automatic flagging is where the '5 – outlier' qualifier flag is applied to all

analytical species from the PTFE and nylon filters (i.e. elements and ions, respectively) when both the sulfur/sulfate ratio and potassium/potassium ion ratios are outside of the predefined limits. This is the only flag that is applied from the automated analytical checks for ions, carbon, and XRF-IC.

Another automated check attempts to identify swaps of sample data. The swap check calculates two indices, one assuming two samples were not swapped and the other assuming the samples were swapped. These indices use the sulfur/sulfate ratio to identify potentially swapped samples for PTFE and nylon filter samples. Similar to the *XRF-IC check* output, the swapped variable will contain concentrations for sulfur and sulfate, outliers by the sulfur/sulfate ratio, and potentially swapped samples based on the calculated indices.

The other functions performed by the *csn_validate* command include sample-field blank swap check (function name: *get_blank_swaps*), a mass check (function name: *mass.check*), a completeness check (function name: *csn_complete*), qualifier flag checks (function name: *qualifiers*), and a function to process invalid filters (function name: *csn_X*). The sample-field blank swap check compares field blanks with sample filters to determine if a filter swap may have occurred. The *mass check* uses the collocated mass concentrations retrieved from AirNowTech to compare reconstructed mass with collocated mass. When available, it also compares gravimetric mass with reconstructed mass. No flags are applied during the mass check; however, the output is useful for the analyst to identify potential issues. The completeness check ensures that all reportable parameters for every sample event have complete records. If any expected records are missing, the function fills in with a "AM - Miscellaneous Void" flag for clarification during review in DART SLT agencies. The qualifier flag check and invalid filters check ensure that flags/null codes are associated with appropriate parameters and output lists of all records with qualifier and null flags, respectively. As part of the qualifier flag check, the Intended Use Date and Sample Start Date of filters are compared. If the dates do not match, the "2 – Operational Deviation" qualifier and an appropriate comment are added to the filter parameters. As part of the invalid filters check, the "AI – Insufficient Data (cannot be calculated)" null code is applied to composite parameters when it is not possible to calculate the concentrations.

Although these checks are performed automatically within *csn_validate*, the analyst should review the outputs from all of the checks to further investigate issues and/or confirm that flags should be applied. To further investigate issues, the analyst may need to work with the sample handling laboratory or measurements laboratories and/or use other available tools, such as regional concentration comparisons, to determine if additional flags or comments should be applied manually for the SLT to validate.

8.3 Review Summary Plots

To generate a summary plot, the analyst can use a variety of R plotting tools or custom plotting tools in the d*atvalCSN* package. The analyst should review the help pages for

each of the available plotting functions. For example, to view the results of the XRF-IC check function at a particular CSN site AA-BBB-CCCC, the analyst can run:

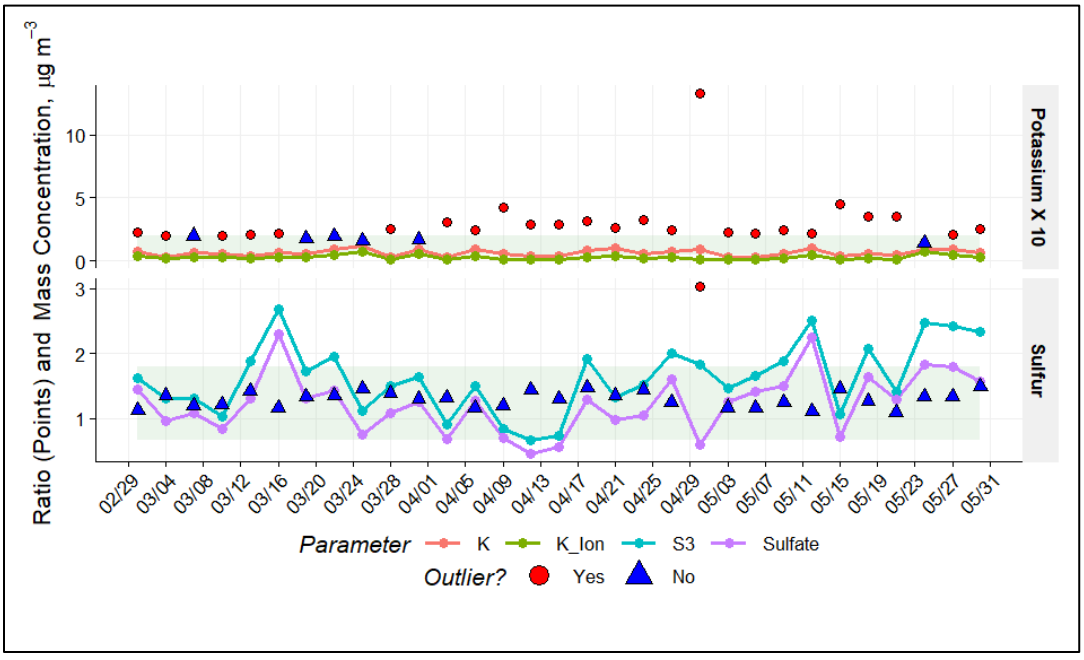
[results] <- datvalCSN::csn_get(start.date = ['YYYY-MM-DD'], end.date = ['YYYY-MM-DD'], Site = ['AA-BBB-CCCC'], FBs = [FALSE])

[xrfic] <- datvalCSN::xrfic.check([results])

datvalCSN::plot_xrfic([xrfic], AqsSiteId =['AA-BBB-CCCC'], POC = ['D'], y.limits = c([y1, y2]))

where "*AA*" is the state code, "*BBB*" is the county code, "*CCCC*" is the site code, and "D" is the POC. Typically, only the sample data is to be plotted so to reduce database querying time, the option to retrieve field blank data, *FBs*, is set to FALSE. If a start and end date is not specified then the number of days back in time from the last sampling date of the processed data available in the database is used. This can be specified by the *csn_get* function parameter *days*. To put the current month of data into context, it is recommended that at least 90 days of data is reviewed. The resulting graph would be similar to Figure 1.

Figure 1. Sample output of *plot_xrfic* for comparing XRF analysis to IC analysis. The turquoise and pink lines are the sulfur and elemental potassium measurements by ED-XRF, the purple and green lines are sulfate and potassium ion measurements by IC, and the points are the ratio of ED-XRF/IC. They also illustrate whether the ratio of the two measurements can be considered an outlier or not.



It may be necessary to change the axis ranges on a plot to view the data more closely in which case the *y.limits* variable can be utilized. However, as data is reviewed on a monthly basis, for convenience, the analyst can enter the following command:

create_monthly_plots(['Month'], [MonthData])

to produce plots for each site and each check. The default directory for saved plots is *U:/CSN/QA/Temporary Plots/*, with a subfolder based on the month of interest (e.g., *Month Plots*). The *['Month']* parameter is a character string that will be used to name the plot folder while the *[MonthData]* parameter is the "validation" object created by the *csn_validate* function. The analyst should cycle through each check and each site to identify potentially swapped samples as well as aberrant data that has not yet been flagged.

8.4 Review all data via web application

In order to facilitate efficient review of all data for a given month, custom web applications (webapps) were developed. Three important webapp tools for the CSN analyst include:

- CSN Data Management – csn.crocker.ucdavis.edu
- CSN Data Explorer – analysis.crocker.ucdavis.edu:3838/csnSites/
- CSN Status Explorer – analysis.crocker.ucdavis.edu:3838/csnStatus/

The CSN Data Management page is an online access point to interact with the UCD CSN database. With this tool, the user can look up filter, batch, and site information. Qualifier and null flags can be assigned here with an explanatory comment. Each time a flag is changed, a timestamped record is made of the change. This is also the portal for importing data from Wood PLC (sample handling laboratory) and DRI (ions and carbon laboratories). Files from the respective folders in the networked U drive are uploaded following the instructions in TI 801A.

The CSN Data Explorer page has multiple tabs for viewing the resultant data in various forms. The analyst should familiarize themselves with the intuitive controls, especially within the "Early Review" (Figure 2), "Explorer" (Figure 3), "Field Blanks" (Figure 4), and "Validation" (Figure 5) tabs. The data displayed in the Early Review tab enables the user to identify anomalies in the PTFE and nylon filters by reviewing the sulfur/sulfate and potassium/potassium ion ratios. It is important to note the "Early Review" tab may be used prior to receiving a complete, monthly dataset. This tool utilizes reported sulfur and sulfate mass loadings before the post processing steps described in TI 801B. The Explorer tab is a comprehensive tool enabling full chemical composition to be assessed using spatial and temporal comparisons. The Field Blanks tab compares field blanks with associated sample filters. This tool enables rapid identification of unusually high field blank mass loadings that may indicate a swap between field blank and sample filters. The Parameter Review tool under the "Validation" tab must be reviewed by the analyst for each monthly dataset. With this tool, every reported data point can be inspected quickly and efficiently within the context of the monthly network data set.

Figure 2. The Early Review tab of the CSN Data Explorer webapp showing the sulfur/sulfate and potassium/potassium ion time series for the specified time range at a specified site.

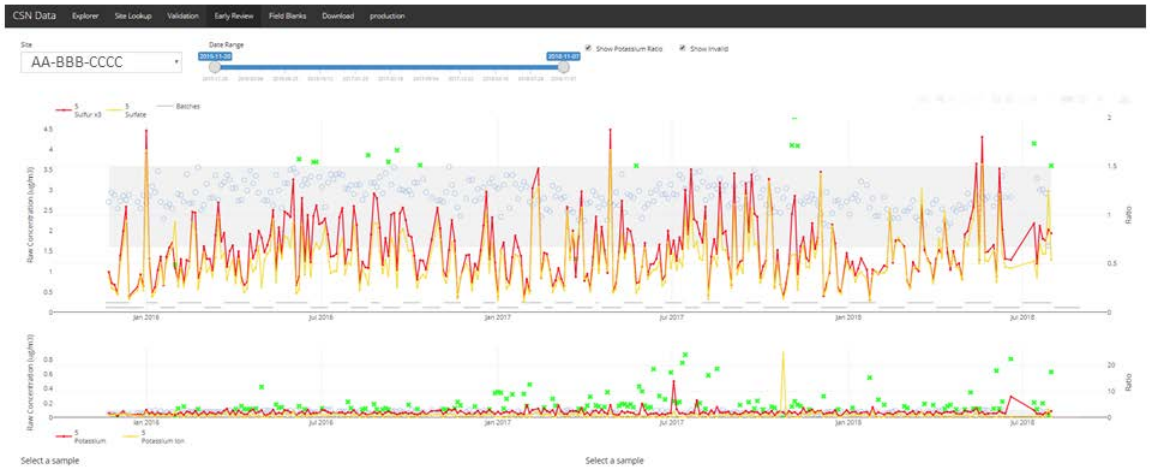


Figure 3. The Explorer tab of the CSN Data Explorer webapp showing various tools available to investigate and compare selected data.

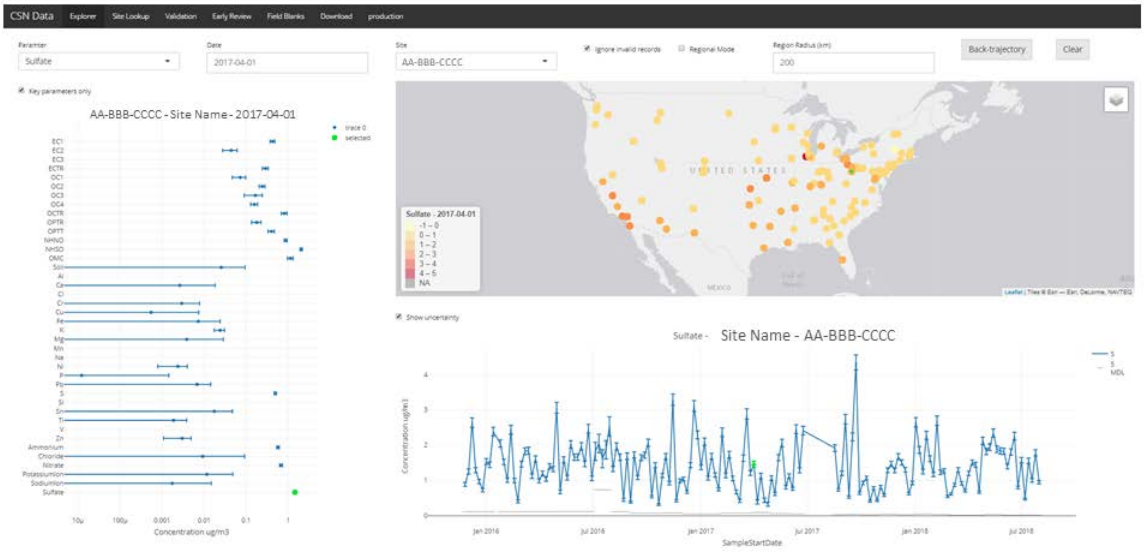


Figure 4. The Field Blanks tab of the CSN Data Explorer webapp showing the mass loading of a selected species from one of the three filter types for all field blank filters across the network.

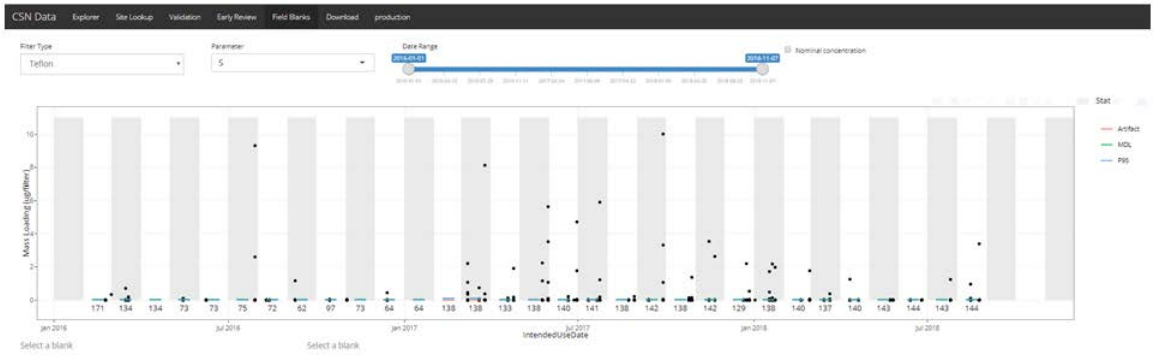
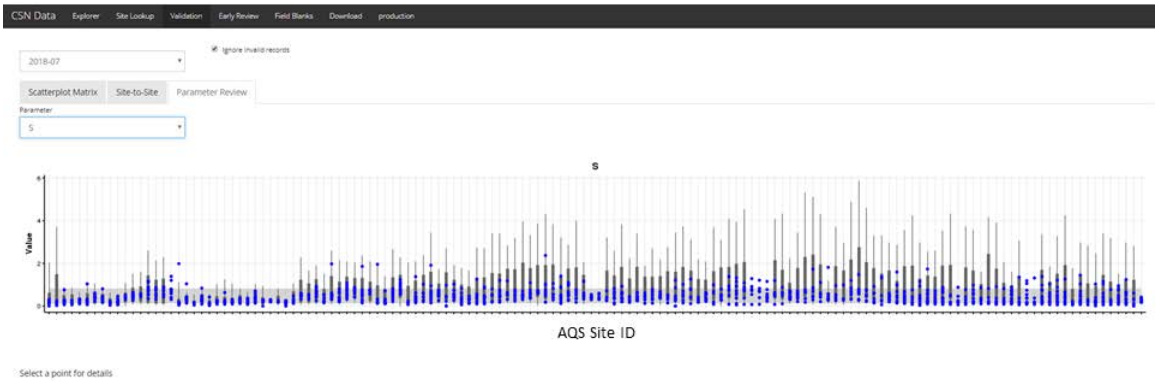
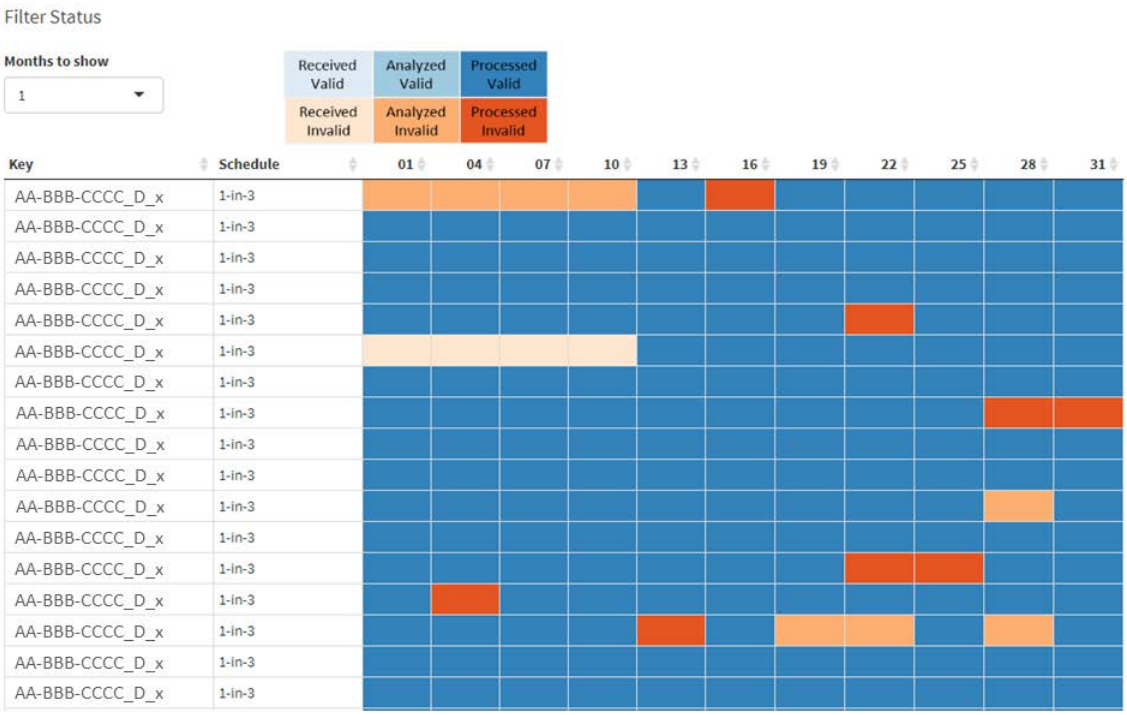


Figure 5. The Parameter Review tool in the Validation tab of the CSN Data Explorer webapp showing the sulfur concentrations for the entire CSN. Sites ordered from west to east on the x-axis and designated by the AQS Site ID. The blue points represent the data for the month selected while the tops of the gray box and whisker indicate historical 90th and 99th percentiles, respectively.



The CSN Status Explorer page is useful for monitoring the progress of analyses and timelines for delivery. There are myriad data views and the user should explore the various figures and tables. An important portion to review for validation is the "Analysis Completeness" tab. Here, the analyst should inspect filter records not analyzed by their respective analyses. All unanalyzed filters should either be flagged as invalid, or be queued for analysis. Similarly, the "Status Grid" is also a useful visual tool to identify consecutive invalid samples, missing analysis data for a given filter record, duplicate issues. This grid can also be used to check overall network issues such as bulk errors in the electronic data. Figure 6 is a screenshot of part of the Status Grid, which uses the data validation webapp.

Figure 6. Snapshot of the Status Grid validation tool on the CSN Status Explorer webapp. The grid is generated for each of the three filter types and shows information for each site from the selected month. Each cell denotes an intended sampling date for a given site and is color coded according to the status of the filter.



Other important views include the “Level 0 Checks” tab, a visual tool of various checks performed dynamically, which can be used to identify issues with the data. If the following criteria are met then the data are shown in the Level 0 Checks as filter records that requiring investigation:

- The start date of a sample filter is missing and there are no comments associated with the filter;
- The end date of a sample filter is missing and there are no comments associated with the filter;
- The start date of a sample filter is not NULL, is the same as the end date, and there are no comments associated with the filter;
- The filter sampled for more than 24 hours and there are no comments associated with the filter;
- The calculated sample volume is more than 10% different from the reported sample volume and there are no comments associated with the filter;
- The sample is marked as invalid but there are no comments associated with the filter;
- The Filter Analysis ID is missing and there are no comments associated with the filter;
- The start date is outside of the expected range for the month of interest;

- The Filter Type does not match the expected Analysis Type;
- The "TT" qualifier flag is inconsistent with the transport temperature reported.

9. DATA VALIDATION EQUATIONS

The following section presents the equations used to calculate swap check indices. These calculations are performed by the *datvalCSN* R package.

The ratio of sulfur to sulfate in the ambient atmosphere is generally well known. Since the majority of sulfur-bearing aerosols by mass are in the form of sulfate, the stoichiometric ratio of 3 x sulfur / sulfate is typically close to one. The calculated indices of the swap check function utilize this tendency. The first index assumes that two subsequent samples are not swapped. The ratio of sulfur by XRF to sulfate by IC is subtracted by one for each sample, and the two results are multiplied. Since both ratios are expected to be close to unity, a small number is expected. Inversely, the second index assumes two samples have been swapped. This irregularity would result in a significantly larger number than index 1. Mathematically,

$$Index1 = \left(\frac{S3_1}{SO4_1} - 1\right) * \left(\frac{S3_2}{SO4_2} - 1\right) \quad (1)$$

$$Index2 = \left(\frac{S3_1}{SO4_2} - 1\right) * \left(\frac{S3_2}{SO4_1} - 1\right) \quad (2)$$

Where,

$S3_x$ = sulfur concentration ($\mu g/m^3$) multiplied by 3

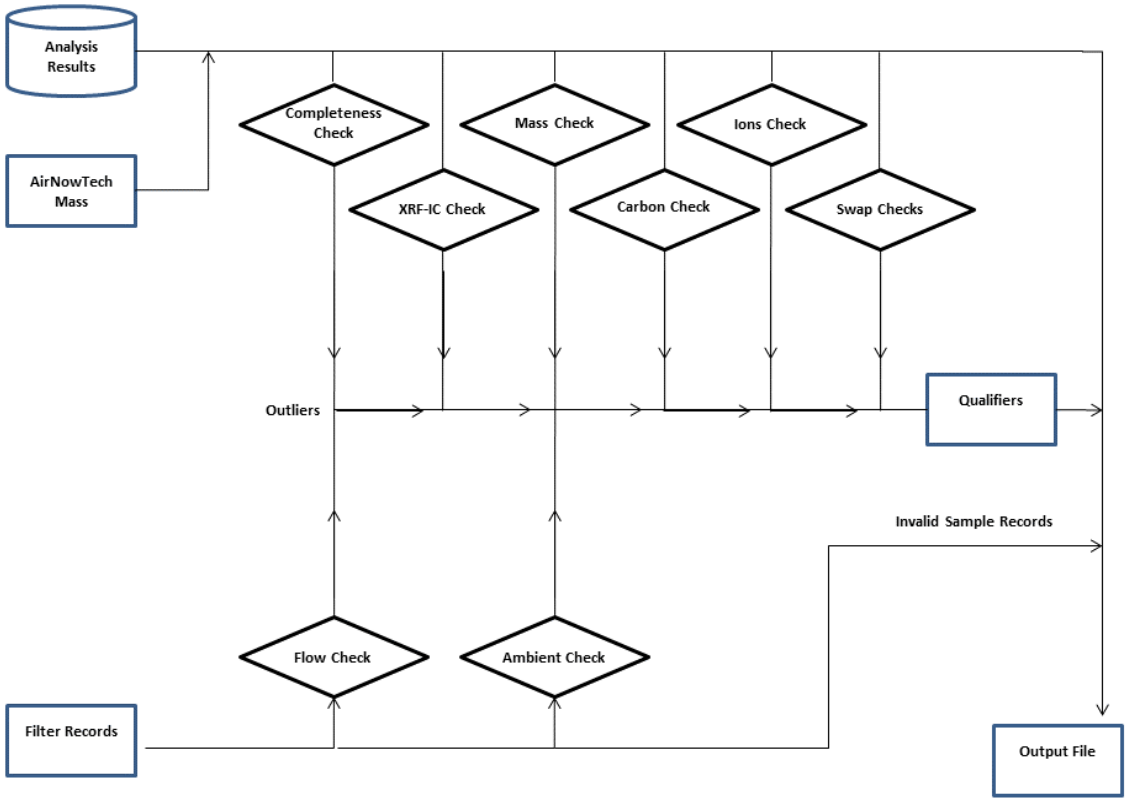
$SO4_x$ = sulfate concentration ($\mu g/m^3$)

Subscripts denote the sampling event of interest (1 for target event, 2 for subsequent event). Empirically, potential swaps are indicated by an Index1 < -0.03 and an absolute value of Index2 < 0.05.

10. DATA PROCESSING CODE

This section describes the data flow through the data validation code used to execute CSN validation checks. Figure 7 outlines the flow of data from the filter and analysis results database tables to final results. The wrapper function *csn_validate* is the only function executed directly by the analyst (see Section 8.1); *csn_validate* in turn calls several functions sequentially to generate data frames with outliers identified. Source code for the functions shown in Figure 7 is stored in the AQRC source repository.

Figure 7. Flow diagram of the validation code in datvalCSN::csn_validate. Rectangles represent data files, diamonds represent R functions, cylinders represent databases, and lines represent inputs and outputs.



11. EQUIPMENT AND SUPPLIES

The associated hardware and software used for CSN data validation are described in the associated UCD SOP #801.

12. QUALITY ASSURANCE AND QUALITY CONTROL

Software bugs and data management issues are tracked through JIRA tracking software. All users have access to our internal JIRA website and can submit, track, and comment on bug reports.

13. REFERENCES

Not Applicable.