

# UCD IMPROVE Standard Operating Procedure #351

## Data Processing and Validation

*Interagency Monitoring of Protected Visual Environments  
Air Quality Research Center  
University of California, Davis*

*Version 4.0*

Updated By: Xiaolu Zhang

Date: 03/14/2019

Approved By: Katrine Gorham

Date: 03/14/2019



## TABLE OF CONTENTS

1. Purpose and Applicability .....	5
2. Responsibilities .....	5
2.1 Data and Reporting Group Manager .....	5
2.2 Quality Assurance Officer .....	5
2.3 Data Validation Specialist .....	5
3. Required equipment and materials .....	6
4. Data Ingestion .....	7
4.1 Carbon Results .....	7
4.2 Ions Results .....	8
4.3 Element and Optical Absorption Results .....	9
4.4 Re-ingesting .....	10
4.5 Issue Tracking .....	10
5. Data Processing .....	10
5.1 Units .....	11
5.2 Artifacts .....	12
5.3 Volume .....	13
5.4 Concentration, Uncertainty, and Method Detection Limit .....	14
5.4.1 PM <sub>2.5</sub> and PM <sub>10</sub> Mass (1A and 4D Modules) .....	15
5.4.2 Ions (2B Module) .....	16
5.4.3 Carbon (3C Module) .....	17
5.4.4 Elements (1A Module) .....	19
5.4.5 Laser Absorption (1A Module) .....	21
5.5 Equations of Composite Variables .....	22
5.5.1 Sulfate by XRF (S3) and Ammonium Sulfate (NH <sub>4</sub> SO <sub>4</sub> ) .....	22
5.5.2 Ammonium Nitrate (NH <sub>4</sub> NO <sub>3</sub> ) .....	23
5.5.3 Soil .....	23
5.5.4 Non-Soil Potassium (KNO <sub>3</sub> ) .....	23
5.5.5 Organic Carbon by Mass (OMC) .....	24
5.5.6 Reconstructed Mass Using Carbon Measurements (RCMC) .....	24
5.5.7 Reconstructed Fine Mass (RCMN) .....	25
6. Data Validation .....	25
6.1 Definition of Status Flags .....	26
6.2 Level 1 Validation Procedures .....	27
6.2.1 Flow Validation .....	28
6.2.2 Level 1B Checks .....	31
6.3 Level 2 Validation Procedures .....	31
6.3.1 Cross-Module Comparison .....	31
1A Module versus 3C Module .....	32
1A Module versus 4D Module .....	33

PM <sub>2.5</sub> reconstructed mass versus gravimetric mass .....	34
6.3.2 Long-Term Network-Wide Checks .....	35
6.3.3 Final Review .....	38
7. Data Delivery .....	39
7.1 Submission to CIRA .....	39
7.2 Submission to AQS.....	39
7.3 Submission to UCD CIA .....	40
8. References.....	40

## LIST OF FIGURES

Figure 1. Carbon analysis results upload page.....	8
Figure 2. Ions analysis results upload page. ....	9
Figure 3. Data processing flow chart. ....	11
Figure 4. PM <sub>2.5</sub> sampler flow rate versus the cyclone pressure transducer reading. ....	30
Figure 5. S <sub>3</sub> /SO <sub>4</sub> comparison plot for the GRGU1 site. ....	32
Figure 6. Comparison plot of light absorption coefficient measurements.....	33
Figure 7. Time series plot of PM <sub>10</sub> and PM <sub>2.5</sub> masses and their ratio at BOWA1 site. ....	34
Figure 8. Time series plot of PM <sub>2.5</sub> gravimetric mass .....	35
Figure 9. Scatter plot of sulfur (×3) versus sulfate for the entire IMPROVE network. ....	36
Figure 10. Scatter plot of chlorine versus chloride for the entire IMPROVE network. ....	36
Figure 11. Scatter plot of fabs versus BC for the whole network.....	37
Figure 12. Ratio of PM <sub>2.5</sub> mass (1A) over PM <sub>10</sub> mass (4D) at ACAD1 site.....	37
Figure 13. Multi-year monthly 10% percentile (top), median (middle) and 90% percentile (bottom) of organic carbon (OC) concentrations.....	38

## LIST OF TABLES

Table 1. Automated validity checks performed during carbon data upload. ....	8
Table 2. Automated validity checks performed during the ions data upload. ....	9
Table 3. Units of data delivered to the CIRA, AQS and UCD CIA databases. ....	12
Table 4. Universal flow constants for the V4 controllers.....	14
Table 5. Fractional uncertainty for the mass.....	16
Table 6. Fractional uncertainty for ions.....	17

Table 7. Analytical detection limits for the carbon species (from DRI).....	18
Table 8. Fractional uncertainty for the carbon species. ....	19
Table 9. Fractional uncertainty for the elemental species.....	20
Table 10. Fractional uncertainty for the laser absorption data.....	22
Table 11. Status flags and their definitions.....	27
Table 12. Definitions and application criteria of automatic flow flags for PM <sub>2.5</sub> .....	29
Table 13. Definitions and application criteria of automatic flow flags for PM <sub>10</sub> . ....	30

## **1. PURPOSE AND APPLICABILITY**

This standard operating procedure (SOP) provides an overview of the procedures for processing and validating the sampling and analytical laboratory data for the IMPROVE network. Data processing and data validation are performed in parallel.

## **2. RESPONSIBILITIES**

This section describes the responsibilities of the individuals involved in data processing and validation.

### **2.1 DATA AND REPORTING GROUP MANAGER**

The Data and Reporting Group Manager oversees all aspects of data ingestion, processing, validation, and reporting.

### **2.2 QUALITY ASSURANCE OFFICER**

The quality assurance officer

- devises techniques that improve the efficiency, traceability, and accuracy of the data management;
- communicates with laboratories regarding analytical issues and/or reanalysis requests;
- develops validation criteria, automated and manual checks, and visualization tools for assessing data quality and consistency;
- reviews measurement detection limit (MDL) and uncertainty;
- analyzes time-series and spatial trends in network data to assess data consistency due to sampling, measurement or procedural changes;
- identifies sampling or measurement deficiencies and proposes solutions/improvements;
- critically evaluates the data using knowledge of air quality and atmospheric chemistry to better understand trends and biases in the data at program level scale.

### **2.3 DATA VALIDATION SPECIALIST**

The data validation specialist

- receives and ingests the analytical data to the University of California, Davis (UCD) IMPROVE database;
- reviews operational and analytical data for errors or incompleteness;
- processes species concentrations and posts monthly dataset to the UCD IMPROVE database;
- performs automated and manual validation checks on concentration data and determines the validity of samples;
- submits Level 2 validated data to project sponsors, Cooperative Institute for Research in the Atmosphere (CIRA), the EPA Air Quality System (AQS), and UCD CSN & IMPROVE Archive (CIA) databases.

### 3. REQUIRED EQUIPMENT AND MATERIALS

The data processing and validation requires all operational and analytical data be loaded into the UCD IMPROVE database (Improve\_2.1). The types of data include:

- Basic filter information such as sample date, site, purpose, and status. These data are recorded during filter preparation and handling and are stored in the filter.Filters table.
- Flow rates. Raw flow readings are either acquired from sampler flashcards (for V2 controllers) or uploaded in real time by the controller (for V4 controllers) and are stored in the sampler.FlowSourceData table (V4 controller data are stored in sampler.FlowSourceDataV2 table). In addition, handwritten log sheets that contain flow readings and other sampling information recorded by the operator are stored in filter.Filters. An SQL procedure called sampler.spFilterAverageFlowRates is used to calculate 24-hour average flow rates for each filter based on the raw flow readings or log sheet data. These are stored in the sampler.AverageFlows table.
- Pre- and post-sampling filter mass values. These are acquired in the UCD Sample Handling Laboratory and stored in the analysis.Mass table.
- Carbon analysis results. These are acquired from files generated by DRI TOR Laboratory and are stored in the analysis.Carbon table.
- Ions analysis results. These are acquired from files generated by RTI IC Laboratory and are stored in the analysis.Ions table.
- Elements analysis results. These are acquired from the UCD XRF Laboratory through a custom ingestion process and are stored in the xrf.MassLoadings table.
- Optical absorption analysis results. These are acquired from the UCD Hybrid Integrating Plate/Sphere (HIPS) Laboratory through a custom ingestion process and are stored in the analysis.HIPS table.

UCD has developed several custom tools for data processing and validation:

**crocker:** This program (a package in the R programming language) provides functions for processing raw filter weights, mass loadings, and flow rates into concentrations, uncertainties, and MDLs. *crocker* also provides utility functions that are used in the online data validation tools (see Section 6).

**datvalIMPROVE:** This R package provides functions for performing routine validation and quality control (QC)(see Section 6.3).

**IMPROVE Management Website:** <http://webapp.improve.crocker.ucdavis.edu/>. This web application provides all UCD laboratory staff with viewing access to relevant tables within the UCD IMPROVE database. Functions within the application with relevance to data processing and validation include the following:

- XRF Section - an interface for processing XRF elemental mass loadings, managing processed sets, and applying flags.
- Filter Section - web pages for searching for specific filters, reviewing operational and analytical data associated with a filter, or applying flags and comments.

- Analysis Data Section - pages for importing and viewing carbon, ions, and optical absorption data, as well as exporting final processed and validated concentration data to AQS format for delivery.
- Operations Section - a live display of the sampler status for the sites equipped with the V4 controllers.

**Flow Graphs:** <http://analysis.crocker.ucdavis.edu:3838/FlowRates/>. This web application provides interactive visualizations of the raw 15 minute flow rates and temperatures as well as the processed 24-hr average flow rate in the UCD IMPROVE database.

**IMPROVE Data Site:** <http://analysis.crocker.ucdavis.edu:3838/ImproveData/>. This web application provides interactive visualizations of processed concentrations, uncertainties, and MDLs, plus custom tools for validation as described in Section 6.3.

## 4. DATA INGESTION

Prior to data processing and validation, data are ingested for each of the analysis pathways: (1) carbon results from DRI, (2) ion results from RTI, and (3) elemental and optical absorption results from UCD.

### 4.1 CARBON RESULTS

Carbon analysis results are sent from DRI to UCD via email in .xml format, including three files:

1. CarbonData.xml
2. CarbonInformation.xml
3. CarbonLaser.xml

All three files are ingested using the UCD IMPROVE Management website. Figure 1 shows a screenshot of the carbon data upload page. To ingest a file from the data upload page, select the file, create a name for the import batch under *Batch Label*, and click submit. CarbonInformation, CarbonLaser, and CarbonData are ingested simultaneously, and an automated validity check is performed (Table 1). Results from the validity check will indicate upload failures. The Data Validation Specialist will review the upload results and notify the Quality Assurance Officer if there are upload failures from validation errors. After ingest, the source files are stored on the file server (U:\IMPROVE\RawDataReceived\Carbon DRI\Imported).

**Figure 1.** Carbon analysis results upload page.

Improve Management Site Home XRF **Analysis Data** Operations Admin Hello Yama Noorzai Log off

Analysis Data Mass **Carbons** Ions HIPS Export Results

### Carbon Data Ingestion

**Carbon Data File:**  
 Choose File No file chosen

**Carbon Information File:**  
 Choose File No file chosen

**Carbon Laser File:**  
 Choose File No file chosen

**Batch Label:**

☐ Ignore warnings  
 (Used mostly for ignoring the warning about duplicate records.)

**Submit**

© 2019 - IMPROVE Data Management Application

**Table 1.** Automated validity checks performed during carbon data upload.

Check	Action
Basic schema validation on xml files	Warning message
No filter is found for record	Warning message
Multiple records for a parameter filter pair	Warning message
Parameter missing for a filter	Warning message
Parameter already recorded in database	Warning message
Import file does not use the same units for each parameter	Warning message
Filters belong to more than one batch	Warning message
No CarbonLaser entry for filter	Warning message

## 4.2 IONS RESULTS

Ions analysis results are sent as one file from RTI to UCD via email in .csv format. The naming convention of the ions data includes the year followed by the ions data set number (eg. ‘2018 1 2 3 data export to UCD’).

The ion analysis records are ingested using the UCD IMPROVE Management website. Figure 2 shows a screenshot of the ions data upload page. To ingest a file form the data upload page,

select the file and click submit. An automated validity check is performed, and results from the validity check will indicate if there are upload failures (Table 2). The Data Validation Specialist will review the upload results and notify the Quality Assurance Officer if there are upload failures from validation errors. After ingest, the source files are stored on the file server (U:\IMPROVE\RawDataReceived\Ions RTI\Ingested).

**Figure 2.** Ions analysis results upload page.

Improve Management Site Home XRF Analysis Data Operations Admin Hello Yama Noorzai Log off

Analysis Data Mass Carbons Ions HIPS Export Results

### Ions Analysis Upload Data

Select Ions analysis source file (.csv file type required):

No file chosen

© 2019 - IMPROVE Data Management Application

**Table 2.** Automated validity checks performed during the ions data upload.

Check	Action
Basic schema validation on xml files	Warning message
No filter is found for record	Warning message
Multiple records for a parameter filter pair	Warning message
Parameter missing for a filter	Warning message
Parameter already recorded in database	Warning message
Import file does not use the same units for each parameter	Warning message
Filters belong to more than one batch	Warning message

### 4.3 ELEMENT AND OPTICAL ABSORPTION RESULTS

Elemental analysis is performed at UCD. The PANalytical XRF software generates results files, which are automatically ingested. The results files are transmitted to a directory on the PANalytical XRF PC (C:\PANalytical\Transmission), and a Windows service (internally named *XRF Data Transfer*) monitors a transmission directory, checking hourly for new files. The XRF results files are standard text files with the extension .qan. The file name includes XRF analysis dates and times in the format YYYYMMDDHHMMSS.qan. The results files and contents are automatically parsed and ingested into tables in the UCD IMPROVE database.

Optical absorption measurements using the Hybrid Integrating Plate/Sphere (HIPS) are performed at UCD. The HIPS instrument generates results in an Excel template file. The data are

then filtered to exclude results not related to samples (e.g. hole measurements), and are copied/pasted into a spreadsheet formatted for upload into the database and saved as a comma separated value (csv) file. The csv file of results are uploaded by a laboratory technician into the UCD IMPROVE database through the IMPROVE web app.

#### **4.4 RE-INGESTING**

If errors are identified requiring new files from DRI or RTI, new files must be provided for ingestion. Upload the new files using the process described in Sections 4.1 and 4.2; the ingest process identifies any changed records. For carbon, take care to ingest with the *ignore warnings* box unchecked and scrutinize the changed records to ensure that they are correct before re-running the process with the *ignore warnings* box checked; only changed records will be overwritten. For ions, the data are ingested without any changes to the original process; the QC code is updated in the UCD IMPROVE database.

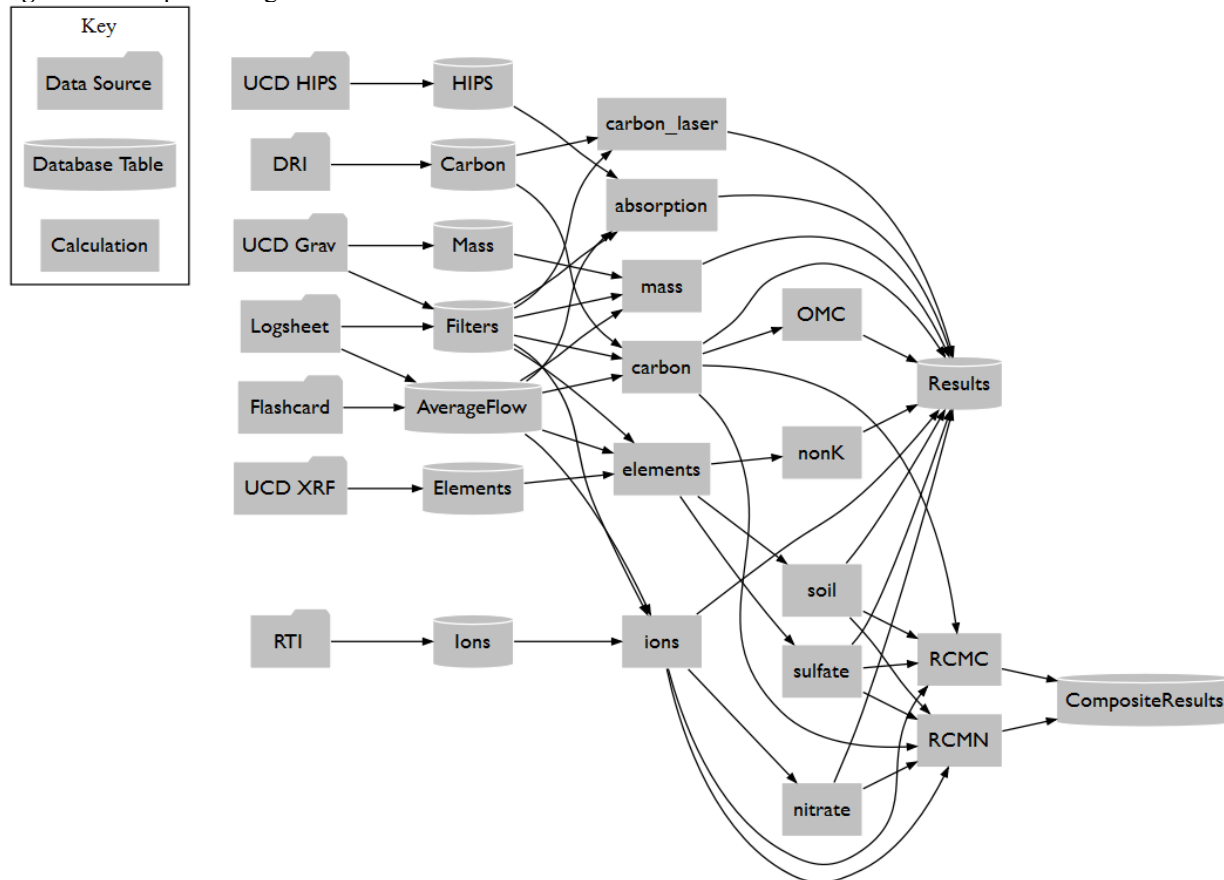
#### **4.5 ISSUE TRACKING**

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.

### **5. DATA PROCESSING**

Data processing for IMPROVE consists of reducing and combining data from the sampling and analytical laboratories to calculate concentrations, uncertainty estimates, and method detection limits (MDLs). Figure 3 shows a flow chart for the IMPROVE data processing.

**Figure 3.** Data processing flow chart.



Calculation of concentrations and associated uncertainties and MDLs are performed within the *crocker* R package, while flow rate calculations are performed in the UCD IMPROVE database. To calculate values for all measured and derived parameters, the following command is run in an R environment:

```
[month_data] <- crocker::improve_calculate_and_post ([YYYY], [MM], server,
AnalysisQcCode = 1, comment = "", replacingId = NULL, replacingQcCode = NULL)
```

This command will calculate concentrations, uncertainties, and MDLs for all measured and derived parameters for the year ([YYYY]) and month ([MM]) and upload the results to the UCD IMPROVE database (*server*) specified in the command in preparation for validation. The processed concentration data are appended to the analysis.Results and analysis.CompositeResults table in the UCD IMPROVE database (Improve\_2.1). A record that contains summary information for the data set, including the comment and the AnalysisQcCode, is inserted into the analysis.Sets table. An AnalysisQcCode of 1 is used for valid routine data.

## 5.1 UNITS

Table 3 lists the data types, parameters, and units for all data submitted to the CIRA, AQS, and UCD CIA databases. For mass, ions, carbon, elements, and light absorption; the units listed are

also used for uncertainty and MDL. NA indicates that the data type is not reported to the corresponding database.

**Table 3.** Units of data delivered to the CIRA, AQS and UCD CIA databases.

Data type	Parameter	CIRA unit	AQS unit	UCD CIA unit
Flow Rate	Flow	L/min	NA	NA
Elapsed Time	ET	hours	NA	NA
Gravimetric mass	PM2.5, PM10	ng/m <sup>3</sup>	µg/m <sup>3</sup>	µg/m <sup>3</sup>
Ions	Cl <sup>-</sup> , NO <sub>2</sub> <sup>-</sup> , NO <sub>3</sub> <sup>-</sup> , SO <sub>4</sub> <sup>2-</sup>	ng/m <sup>3</sup>	µg/m <sup>3</sup>	µg/m <sup>3</sup>
Carbon	OC1, OC2, OC3, OC4, OC, OPTR, EC1, EC2, EC3, EC <sub>4</sub>	ng/m <sup>3</sup>	µg/m <sup>3</sup>	µg/m <sup>3</sup>
	TC, OPTT, OPTR at other wavelength, OPTT at other wavelength	ng/m <sup>3</sup>	NA	NA
Carbon_laser	RefF_wavelength, Refl_wavelength, RefM_wavelength, TransF_wavelength, Transl_wavelength, TransM_wavelength	reading	NA	NA
Elements	Na, Mg, Al, Si, P, S, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Pb, Se, Br, Rb, Sr, Zr	ng/m <sup>3</sup>	µg/m <sup>3</sup>	µg/m <sup>3</sup>
Light absorption	f <sub>abs</sub>	Mm <sup>-1</sup>	NA	NA
Composite species	OMC, NHNO, NHSO, PM <sub>10</sub> -PM <sub>2.5</sub>	NA	µg/m <sup>3</sup>	µg/m <sup>3</sup>

## 5.2 ARTIFACTS

An artifact is defined as any increase or decrease of material on the filter that positively or negatively biases the measurement of ambient concentration. Artifact corrections are applied to the ion, carbon, and element measurements. Artifact examples include:

- (1) Contamination of the filter medium (positive).
- (2) Contamination acquired by contact with the cassettes or in handling (positive).
- (3) Adsorption of gases during collection that are erroneously measured as particles (positive).
- (4) Volatilization of particles during collection and in handling (negative).
- (5) Fall-off of particles during handling after collection (negative).

For the ion measurements, the artifact correction method attempts to account for the first two types of artifacts, which is estimated using data from field blanks. The field blanks are handled as normal filters (loaded into cassettes and cartridges, shipped to and from the field, and left in the sampler for a week) except that no air is drawn through them. The field blanks are collected randomly at all sites on a periodic basis. When there are  $\geq 50$  field blanks in a month, the artifact corrections are calculated as the median loading measured on the field blanks. Otherwise, values from the previous month(s) are included until at least 50 field blanks are available. Artifact corrections are subtracted from each ambient concentration for the corresponding month.

For the carbon measurements, the artifact correction method attempts to account for the first three types of artifacts, and are estimated using data from field blanks (beginning with the 2005 data). The field blanks are handled as normal filters (loaded into cassettes and cartridges, shipped to and from the field, and left in the sampler for a week) except that no air is drawn through

them. The field blanks are collected randomly at all sites on a periodic basis. When there are  $\geq 50$  field blanks in a month, the artifact corrections are calculated as the median loading measured on the field blanks; otherwise, values from the previous month(s) are included until at least 50 field blanks are available. Artifact corrections are subtracted from each ambient concentration for the corresponding month. For further background information and detail regarding past use of stacked filters for artifact correction and subsequent application of a correction factor, see data advisory: [http://vista.cira.colostate.edu/Improve/wp-content/uploads/2016/04/Dillner\\_OCArtifactAdjustmentIMPROVEOct2012.pdf](http://vista.cira.colostate.edu/Improve/wp-content/uploads/2016/04/Dillner_OCArtifactAdjustmentIMPROVEOct2012.pdf) and [http://vista.cira.colostate.edu/improve/Data/QA\\_QC/Advisory/da0032/da0032\\_OC\\_artifact.pdf](http://vista.cira.colostate.edu/improve/Data/QA_QC/Advisory/da0032/da0032_OC_artifact.pdf)

Measurements are not corrected for the two negative artifact types (volatilization and fall-off). The measured mass loadings for the higher-volatility organics may be much less than those in the atmosphere because of volatilization of particles during the remainder of the sampling or during transportation. Volatilization of nitrate and chloride from the nylon filters is assumed to be insignificant. Depending on the environmental conditions, some ammonium nitrate collected on polytetrafluoroethylene (PTFE) filters may volatilize. In those cases, the fine mass on the PTFE filter may underestimate the ambient PM<sub>2.5</sub> mass concentrations.

For discussion of artifact correction for element measurements, see Section 5.4.4.

### 5.3 VOLUME

The sample volume is a product of the flow rate and the sampling duration. The sampling duration is determined using elapsed time (ET) as recorded by the sampler controller.

For the PM<sub>2.5</sub> modules (1A, 2B, and 3C modules), the flow rate is determined from the measurement of the pressure drop across the cyclone using a pressure transducer. Prior to 2016, the 15-minute pressure measurements were averaged over the whole sampling period (nominally 24 hours) for calculating the average flow rate. Beginning with January 2016 data, the average flow rate is an elapsed time-weighted average, calculated from the individual 15-minute pressure measurement. The sampler flow rate for 1A, 2B, and 3C modules is calculated using equation 351-1.

$$Q = 10^a M^b * F(elev) * \sqrt{\frac{T + 273.15}{293.15}} \quad (351-1)$$

Q = volumetric flow rate (using site-specific temperature and pressure, not STP)

a, b = calibration coefficients

M = cyclone transducer reading. If the transducer readings are taken from the controller screen, they can be used in equation 351-1 directly. If the transducer readings are taken from the flashcard file, they must be divided by 100.

$F(elev)$  = elevation factor to account for pressure difference between sea level and site.

T = ambient temperature in degrees Celsius at time of sampling.

For the PM<sub>10</sub> module (4D module), the sampler flow rate is calculated using equation 351-2.

$$Q = (c + d * G) * F(elev)^2 * \sqrt{\frac{T + 273.15}{293.15}} \quad (351-2)$$

Q = volumetric flow rate

c, d = calibration coefficients

G = critical orifice transducer reading. If the transducer readings are taken from the controller screen, they can be used in equation 351-2 directly. If the transducer readings are taken from the flashcard file, they must be divided by 100.

$F(elev)$  = elevation factor to account for pressure difference between sea level and the site.

T = ambient temperature in degrees Celsius at time of sampling.

The calibration coefficients (a, b, c, and d) in equations (351-1) and (351-2) have historically been site-specific. Starting with the 1/1/2018 samples onward, a set of universal flow constants for the V4 controller cyclone and orifice channels are determined using the flow data collected from the previous year and will be updated annually, although the values are not expected to vary significantly from year to year (Table 4).

**Table 4.** Universal flow constants for the V4 controllers.

Module	Intercept (a, c) (reported for data 1/1/2018-current for sites with V4 controller)	Slope (b, d) (reported for data 1/1/2018-current for sites with V4 controller)
PM <sub>2.5</sub>	1.491	0.3839
PM <sub>10</sub>	1.291	1.333

## 5.4 CONCENTRATION, UNCERTAINTY, AND METHOD DETECTION LIMIT

The calculations described in this section are performed in R using the R function listed at the beginning of Section 5.

The concentration is calculated using equation 351-3, where mass of material on the filter is equal to the difference between the mass measured on the sample and the mass on the unused filter. For gravimetric analysis, the mass on the unused filter is determined from the pre-weight of individual PTFE filters. For measurement of ions and carbon, the mass on the unused filter is determined from the median of field blank loadings. For calculation of element concentrations, see Section 5.4.4.

$$C = \frac{A - B}{V} \quad (351-3)$$

C = ambient concentration (ng/m<sup>3</sup>)

A = mass measured on sample (ng/filter or ng/cm<sup>2</sup>)

B = artifact mass (ng/filter) = pre-weight, monthly median of ion or carbon field blanks (FB). Prior to June 2002, artifacts were calculated by seasonal quarter instead of by month.

V = sample air volume (m<sup>3</sup>) = Q \* Elapsed Time

The uncertainty is reported with each concentration. The general model for the uncertainty is a quadratic sum of two components of uncertainty as shown in Equation 351-4.

$$\sigma(c) = \sqrt{[fC]^2 + \left[\frac{\sigma_a}{V}\right]^2} \quad (351-4)$$

$\sigma_a$  = analytical uncertainty. This is a constant term from additive sources of uncertainty, such as those related to background contamination of the filters. For large concentrations, this is small compared to the fractional term.

f = fractional uncertainty. This term results from various sources of proportional uncertainties, such as the analytical calibration and flow rate measurements. Beginning with data from 1/1/2018, fractional uncertainties (f) are determined using the most recent two years of data from collocated measurements (352-5 and 351-6). If the count of collocated pairs over the two year period is less than 60, a value of 0.25 is adopted as f.

$$srd = \frac{(Collo - Routine) / \sqrt{2}}{(Collo + Routine) / 2} \quad (351-5)$$

$$f = \frac{(84th \text{ percentile of } srd) - (16th \text{ percentile of } srd)}{2} \quad (351-6)$$

The MDLs are also reported with each concentration. Beginning with data from 1/1/2018, MDLs for ions, carbon, and elements are calculated as 95<sup>th</sup> percentile minus median of field blanks.

#### 5.4.1 PM<sub>2.5</sub> and PM<sub>10</sub> Mass (1A and 4D Modules)

PM<sub>2.5</sub> mass is measured gravimetrically on the polytetrafluoroethylene (PTFE) filter from the 1A Module. PM<sub>10</sub> mass is measured gravimetrically on the PTFE filter from the 4D Module. The pre- and post-weights (as micrograms per filter) are stored in the analysis.Mass table in the UCD IMPROVE database.

The constant analytical uncertainty,  $\sigma_a$ , in equation 351-4 is equal to 5 µg for all filters. The mass concentration, uncertainty, and MDL in nanograms per cubic meter are calculated using the following equations:

$$C_{Mass} = 10^6 \frac{ng}{mg} * \left( \frac{Postweight - preweight}{V} \right) \quad (351-7)$$

$$\sigma_{Mass} = 1000 \frac{ng}{\mu g} * \sqrt{\left( \frac{5 \mu g}{V} \right)^2 + \left( \frac{C_{Mass} * f}{1000 \frac{ng}{\mu g}} \right)^2} \quad (351-8)$$

$$mdl_{Mass} = 1000 \frac{ng}{\mu g} * \frac{10 \mu g}{V} \quad (351-9)$$

f = fractional uncertainty (Table 5).

**Table 5.** Fractional uncertainty for the mass.

Species	f reported for data 2/28/1995 – 12/31/2006	f reported for data 1/1/2017 – 12/31/2017	f reported for data 1/1/2018 – current
PM <sub>2.5</sub>	0.03	0.03	0.04
PM <sub>10</sub>	0.03	0.07	0.07

#### 5.4.2 Ions (2B Module)

Ions are measured by ion chromatography using the nylon filter from the 2B Module. Ions data (as micrograms per filter) are stored in the analysis.Ions table in the UCD IMPROVE database.

For chloride, nitrate, and sulfate; the concentration, uncertainty, and MDL in nanograms per cubic meter are calculated using the following equations:

$$C_{ion} = 1000 \frac{ng}{\mu g} * \frac{(A_{ion} - B_{ion})}{V_{B module}} \quad (351-10)$$

$$\sigma_{ion} = 1000 \frac{ng}{\mu g} * \frac{\sqrt{(Max(\sigma_{dfb}, mdl_{analytical}))^2 + (f * (A_{ion} - B_{ion}))^2}}{V_{B module}} \quad (351-11)$$

$$mdl = 1000 \frac{ng}{\mu g} * \frac{Max((95th \text{ percentile of } FB - B_{ion}), mdl_{analytical})}{V_{B module}} \quad (351-12)$$

For nitrite, the concentration and MDL equations are the same as above, while the uncertainty equation is affected by the concentration value:

When  $C_{ion} > 0$

$$\sigma_{ion} = 1000 \frac{ng}{\mu g} * \frac{\sqrt{(Max(\sigma_{dfb}, mdl_{analytical}))^2 + (f * (A_{ion} - B_{ion}))^2}}{V_{B module}} \quad (351-13)$$

When  $C_{ion} \leq 0$

$$\sigma_{ion} = 0 \quad (351-14)$$

$B_{ion}$  = median of the field blank measurements when there are  $\geq 3$  field blanks in a month; otherwise, values from the previous month are used.

$mdl_{analytical}$  = 0.03 for chloride, 0.01 for nitrite, 0.05 for nitrate, and 0.07 for sulfate.. The analytical MDL for each ion is to be used as the overall MDL in the event that the median value of the field blanks is lower than the respective analytical MDL.

$\sigma_{dfb}$  = standard deviation of the field blank measurements for the appropriate ion.

f = fractional uncertainty (Table 6).

**Table 6.** Fractional uncertainty for ions.

Species	f reported for data 1/1/2005 – 12/31/2016	f reported for data 1/1/2017 – 12/31/2017	f reported for data 1/1/2018 – current
Chloride (Cl <sup>-</sup> )	0.08	0.08	0.08
Nitrite (NO <sub>2</sub> <sup>-</sup> )	0.22	0.25	0.25
Nitrate (NO <sub>3</sub> <sup>-</sup> )	0.04	0.03	0.04
Sulfate (SO <sub>4</sub> <sup>=</sup> )	0.02	0.02	0.02

### 5.4.3 Carbon (3C Module)

Carbon is measured by thermal optical reflectance (TOR) and thermal optical transmittance (TOT) using the quartz filter from the 3C Module. The seven carbon su2B fractions (as micrograms per filter), together with pyrolyzed carbon (OP) determined by both TOR and TOT (as micrograms per filter), are stored in the analysis. Carbon table in the UCD IMPROVE database. For the carbon su2B fractions (OC1-OC4, EC1-EC3), the primary factors that determine the fractional uncertainty are the homogeneity of the sample deposit and the accuracy of the temperature set point in each stage. For OP fractions, the primary factors that determine the fractional uncertainty are the laser signal stability and the accuracy of the split point placement.

The TOR elemental carbon (ECTR) component is assumed to be all carbon evolved at 580°C and above, after the laser indicates that reflectance has returned to the initial value. The TOR organic carbon (OCTR) component is assumed to be all carbon evolved at 580°C and below, in a pure helium environment, plus the pyrolyzed organic fraction. The total carbon (TC) is sum of OCTR and ECTR. Only the TOR OC and EC are calculated and reported.

The concentration, uncertainty, and MDL in nanograms per cubic meter for the carbon species (OC1, OC2, OC3, OC4, OPTR, OPTT, EC1, EC2, EC3, as well as OCTR, ECTR, TC) are calculated using the following equations:

$$C = 1000 \frac{ng}{\mu g} * \frac{(A - B)}{V_{C module}} \quad (351-15)$$

$$\sigma = 1000 \frac{ng}{\mu g} * \frac{\sqrt{(Max(\sigma_{dfb}, t))^2 + (f * (A - B))^2}}{V_{C module}} \quad (351-16)$$

$$mdl = 1000 \frac{ng}{\mu g} * \frac{Max((95th \text{ percentile of } FB - B), t)}{V_{C module}} \quad (351-17)$$

$B$  = median of the field blank measurements when there are  $\geq 3$  field blanks in that month, otherwise the number from the previous month is used.

$\sigma_{dfb}$  = standard deviation of the field blank measurements for the appropriate carbon species.

$t$  = analytical detection limits (Table 7).

$f$  = fractional uncertainty (Table 8).

**Table 7.** Analytical detection limits for the carbon species (from DRI).

Species	t
OC1	0.51
OC2	0.51
OC3	0.51
OC4	0.51
OPTR, OPTR at other wavelength	0.15
OPTT , OPTT at other wavelength	0.15
EC1	0.15
EC2	0.15
EC3	0.15
ECTR	0.15
OCTR	0.51
TC	0.57

**Table 8.** Fractional uncertainty for the carbon species.

Species	f reported for data 1/1/2005 – 12/31/2016	f reported for data 1/1/2017 – 12/31/2017	f reported for data 1/1/2018 –
OC1	0.23	0.27	0.23
OC2	0.15	0.13	0.11
OC3	0.13	0.13	0.13
OC4	0.15	0.13	0.13
OPTR, OPTR at other wavelength	0.13	0.16	0.20
OPTR, OPTR at other wavelength	0.13	0.12	0.14
EC1	0.10	0.10	0.11
EC2	0.17	0.18	0.19
EC3	0.42	0.25	0.25
ECTR	0.12	0.14	0.14
OCTR	0.08	0.09	0.08
TC	0.08	0.08	0.07

#### 5.4.4 Elements (1A Module)

Elements are measured using X-ray fluorescence (XRF; PANalytical Epsilon 5) using the PTFE filters from the 1A Module since 2011.

The PANalytical XRF instruments report the elements in terms of counts per mV per second, which must be converted into areal densities using element calibration factors (stored in the UCD IMPROVE database). Blank subtraction is performed on the XRF measurements by subtracting the median field blank count from the same instrument that was used to analyze the sample. The field blank correction is specific to each instrument because the counts vary from instrument to instrument. Since the number of field blanks analyzed on a specific instrument during a month may not be statistically sufficient, field blank selection is not grouped by month. Rather, the UCD IMPROVE database uses a procedure (xrf.spCreateFeldBlankSet) to select the last 35 field blanks analyzed before the determination date, which is the date of the last sample analyzed from a one-month batch. The selected 35 field blanks are used to calculate batch and instrument specific blank correction.

$$\sigma = 1000 \frac{ng}{\mu g} * \sqrt{(0.608 * \text{Max}((\text{blankStat.Percentile95} - \text{blankStat.Median}) * \text{p.EValue}, y.DL))^2 + (f * A_{\text{element}})^2} \quad (351-18)$$

$A_{\text{element}} = (\text{netCounts} - \text{blankStat.Median}) * \text{p.EValue}$ , which is the areal density calculated for the element measured by XRF.

$M_{\text{element}} = \max(((\text{blankStat.Percentile95} - \text{blankStat.Median}) * p.EValue), y.DL[(\mu\text{g}/\text{cm}^2)]) * 1000$ , which is the areal MDL reported for the element measured by XRF.

$p.EValue$  = element calibration factor.

$y.DL$  = static MDL value previously used.

$f$  = fractional uncertainty (Table 9).

blankStat values are specific to a batch of 35 filters analyzed on a specific instrument during the time period that the current sample month was analyzed.

**Table 9.** Fractional uncertainty for the elemental species.

Species	f reported for data 1/1/2005 – 12/31/2016	f reported for data 1/1/2017 – 12/31/2017	f reported for data 1/1/2018 – current
Al	0.09	0.08	0.08
As	0.25	0.21	0.25
Br	0.10	0.11	0.10
Ca	0.06	0.07	0.06
Cl	0.14	0.18	0.14
Cr	0.22	0.17	0.15
Cu	0.12	0.11	0.13
Fe	0.06	0.06	0.05
K	0.03	0.05	0.03
Mg	0.15	0.16	0.15
Mn	0.13	0.13	0.14
Na	0.14	0.15	0.14
Ni	0.16	0.16	0.13
P	0.25	0.33	0.27
Pb	0.13	0.13	0.14
Rb	0.25	0.25	0.25
S	0.03	0.03	0.02
Se	0.25	0.12	0.25
Si	0.10	0.07	0.06
Sr	0.16	0.14	0.13
Ti	0.11	0.09	0.09
V	0.12	0.14	0.17
Zn	0.06	0.08	0.08

Zr	0.25	0.25	0.25
----	------	------	------

The areal densities, areal uncertainty, and areal MDL (in units of mass/area) are stored in the UCD IMPROVE database and used to calculate concentrations using the following equations:

$$C_{element} = \frac{A_{element} * (Deposit\ area)}{V} \quad (351-19)$$

$$\sigma_{element} = \frac{U_{element} * (Deposit\ area)}{V} \quad (351-20)$$

$$mdl_{element} = \frac{M_{element} * (Deposit\ area)}{V} \quad (351-21)$$

$A_{element}$  = areal concentration of the element measured by XRF.

Deposit area = area of sample deposit on the filter (cm<sup>2</sup>), determined from the filter holder or mask size.

#### 5.4.5 Laser Absorption (1A Module)

Optical absorption is measured by a hybrid integrating plate and sphere (HIPS) system using the PTFE filter from the 1A Module. The laser absorption measurements are stored as reflectance (R) and transmittance (T) values in analysis.HIPS table in the UCD IMPROVE database. The filter absorption coefficient ( $f_{abs}$ ) values (Mm<sup>-1</sup>) are calculated using the following equation:

$$f_{abs} = \left( 10000 \times \log \left( \max \left( \frac{intercept - slope \times R}{T}, 0.1 \right) \right) \times \frac{Deposit\ Area}{V} \right) \times \frac{1}{100} \quad (351-22)$$

The intercept and slope constants are determined by constructing a calibration curve using field blank data. New calibration constants are obtained after major instrument reconfiguration and/or change of PTFE filter lot.

$$\sigma_{FABS} = \sqrt{\left( \frac{(10000 * 0.014)}{\left( \frac{V}{Deposit\ area} * 100 \right)} \right)^2 + (f * LRNC)^2} \quad (351-23)$$

$$mdl = 10000 * 0.032 * \left( \frac{Deposit\ area}{V} \right) * \frac{1}{100} \quad (351-24)$$

$f$  = fractional uncertainty (Table 10).

**Table 10.** Fractional uncertainty for the laser absorption data.

Species	$f$ reported for data 2/28/1995 – 12/31/2006	$f$ reported for data 1/1/2017 – 12/31/2017	$f$ reported for data 1/1/2018 – current
$f_{abs}$	0.03	0.08	0.06

## 5.5 EQUATIONS OF COMPOSITE VARIABLES

The following composite variables are combinations of the measured concentrations of particles collected on the polytetrafluoroethylene (PTFE) filter from the 1A Module, and are used in the Level 2 validation procedures described in Section 6.3. For the composite variables, concentration is determined along with the uncertainty and MDL. The uncertainty calculations assume that the component concentrations are independent and the multiplicative factors have no uncertainty. The independence assumption is not strictly valid for many composites because of common factors, such as volume. However, the effect on the overall uncertainty is too small to warrant more complicated calculations.

### 5.5.1 Sulfate by XRF (S3) and Ammonium Sulfate (NHSO)

Sulfur is predominantly present as sulfate in the atmosphere. To compare the sulfur by XRF and the sulfate by ion chromatography, the XRF concentration is multiplied by the ratio of sulfate to sulfur atomic mass ( $96.06/32.06 = 3.0$ ). This composite is labeled S3 in the data validation plots.

The sulfate is generally present as ammonium sulfate,  $(NH_4)_2SO_4$ , although it can be present as ammonium bisulfate,  $(NH_4)HSO_4$ , sulfuric acid,  $H_2SO_4$ , gypsum,  $CaSO_4 \cdot 2H_2O$ , and, in marine areas, as sodium sulfate,  $Na_2SO_4$ . In many cases, the particle will include associated water, this is omitted from the calculation. In order to simplify the calculation, all sulfur is assumed to be present as ammonium sulfate. The concentrations, uncertainties, and MDLs are calculated using the following equations:

$$NHSO = 4.125 * S$$

$$S3 = 3 * S \quad (351-25)$$

$$\sigma_{NHSO} = 4.125 * \sigma(S)$$

$$\sigma_{S3} = 3 * \sigma(S) \quad (351-26)$$

$$mdl(NHSO) = 4.125 * mdl(S)$$

$$mdl(S3) = 3 * mdl(S) \quad (351-27)$$

For ammonium bisulfate, sulfuric acid, and sodium sulfate the factors are 3.59, 3.06, and 4.43, respectively. In the first two cases, the actual dry mass associated with sulfate is less than NHSO, and in the third case, more.

### 5.5.2 Ammonium Nitrate (NHNO)

This composite is the total dry concentration associated with nitrate, assuming 100% neutralization by ammonium. The concentrations, uncertainties, and MDLs are calculated using the following equations:

$$NHNO = 1.29 * NO_3^- \quad (351-28)$$

$$\sigma_{NHNO} = 1.29 * \sigma(NO_3^-) \quad (351-29)$$

$$mdl(NHNO) = 1.29 * mdl(NO_3^-) \quad (351-30)$$

### 5.5.3 Soil

The soil component consists of the sum of the predominantly soil elements measured by XRF, multiplied by a coefficient to account for oxygen for the normal oxide forms ( $Al_2O_3$ ,  $SiO_2$ ,  $CaO$ ,  $K_2O$ ,  $FeO$ ,  $Fe_2O_3$ ,  $TiO_2$ ), and augmented by a factor to account for other compounds not included in the calculation, such as  $MgO$ ,  $Na_2O$ , water, and  $CO_2$ . The following assumptions are made:

- Fe is split equally between  $FeO$  (oxide factor of 1.29) and  $Fe_2O_3$  (oxide factor of 1.43), giving an overall Fe oxide factor of 1.36.
- Fine K has a non-soil component from smoke. Based on the K/Fe ratio for average sediment (*Handbook of Chemistry and Physics*),  $0.6 * Fe$  is used as a surrogate for soil K. The oxide factor for K  $\left( K_2O, \frac{39.1 * 2 + 16.0 \text{ g/mol}}{39.1 * 2 \text{ g/mol}} = 1.2 \right)$  is added for a total Fe factor of  $0.72 * Fe$  ( $0.6 * 1.2$ ) for the potassium oxide in soil. This increases the factor for Fe from 1.36 to 2.08.
- The oxide forms of the soil elements account for 86% of average sediment; in order to obtain the total mass associated with soil, the final factors are divided by 0.86 (*Handbook of Chemistry and Physics*). The concentrations, uncertainties, and MDLs are calculated using the following equations:

$$SOIL = 2.2 * \max(Al,0) + 2.49 * \max(Si,0) + 1.63 * \max(Ca,0) + 2.42 * \max(Fe,0) + 1.94 * \max(Ti,0) \quad (351-31)$$

$$\sigma(SOIL) = \sqrt{(2.2 * \max(\sigma(Al),0))^2 + (2.49 * \max(\sigma(Si),0))^2 + (1.63 * \max(\sigma(Ca),0))^2 + (2.42 * \max(\sigma(Fe),0))^2 + (1.94 * \max(\sigma(Ti),0))^2} \quad (351-32)$$

$$mdl(SOIL) = 0 \quad (351-33)$$

The soil variable is calculated for all valid XRF analyses.

### 5.5.4 Non-Soil Potassium (KNON)

Non-soil potassium is the measured fine potassium minus the soil potassium estimated from iron. Non-soil potassium is a qualitative tracer of smoke. However, the ratio of potassium/smoke mass may change as the aerosol ages. Particulate smoke potassium may be produced by the transformation of volatilized potassium, and appears to be in a smaller

size range than most smoke mass. Close to the smoke source, the particulate potassium may not have time to form. For long-range transport, most other smoke mass may settle out more than potassium mass. The concentrations, uncertainties, and MDLs are calculated using the following equations:

$$KNON = (K - 0.6 * FE) \quad (351-34)$$

$$\sigma(KNON) = \sqrt{\sigma^2(K) + [0.6 * \sigma(FE)]^2} \quad (351-35)$$

$$mdl(KNON) = 0 \quad (351-36)$$

The soil factor of 0.6 may vary slightly with the site; this will produce a small positive or negative offset for baseline values when no smoke is present. Therefore, negative values are retained. KNON is calculated for all valid XRF analyses. If a concentration is less than the MDL, the concentration and uncertainty are assumed to be equal to the MDL.

### 5.5.5 Organic Carbon by Mass (OMC)

To determine the total amount of organic mass associated with the organic carbon, the ratio of organic mass to organic carbon is assumed to be 1.8. The concentrations, uncertainties, and MDLs are calculated using the following equations:

$$OMC = 1.8 \times OC \quad OMC = 1.8 \times OC = 1.8 \times (O1 + O2 + O3 + O4 + OP) \quad (351-37)$$

$$\sigma_{OMC} = 1.8 \times \sigma_{OC} = 1.8 \times 1000 \frac{ng}{\mu g} * \frac{\sqrt{(\text{Max}(\sigma_{dfb}, t))^2 + (f * (A_{OC} - B_{OC}))^2}}{V_{Cmodule}} \quad (351-38)$$

$$mdl_{OMC} = 1.8 \times mdl_{OC} = 1.8 \times 1000 \frac{ng}{\mu g} * 2 * \frac{\text{Max}(\sigma_{dfb}, t)}{V_{Cmodule}} \quad (351-39)$$

### 5.5.6 Reconstructed Mass Using Carbon Measurements (RCMC)

The reconstructed mass is the sum of sulfate, soil, salt, elemental carbon, and organic mass. The only components not included are water and nitrate. The concentrations, uncertainties, and MDLs are calculated using the following equations:

$$RCMC = \text{max}(NH_4SO_4, 0) + \text{max}(SOIL, 0) + 1.8 * \text{max}(Cl^-, 0) + \text{max}(LAC, 0) + \text{max}(OMC, 0) \quad (351-40)$$

$$\sigma(RCMC) = \sqrt{(\text{max}(\sigma(NH_4SO_4), 0))^2 + (\text{max}(\sigma(SOIL), 0))^2 + [1.8 * \text{max}(\sigma(Cl^-), 0)]^2 + (\text{max}(\sigma(LAC), 0))^2 + (\text{max}(\sigma(OMC), 0))^2} \quad (351-41)$$

$$mdl(RCMC) = 0 \quad (351-42)$$

RCMC is always positive.

RCMC is more relevant at sites where the neutralization of sulfate may be less than 100%, at sites with high nitrate, and at marine sites.

### 5.5.7 Reconstructed Fine Mass (RCMN)

At sites where ammonium nitrate (NHNO) is present, adding ammonium nitrate to the RCMC can make the reconstructed mass very close to the measured value. The concentrations, uncertainties, and MDLs are calculated using the following equations:

$$RCMN = \max(NHSO, 0) + \max(NHNO, 0) + \max(SOIL, 0) + 1.8 * \max(Cl^-, 0) + \max(LAC, 0) + \max(OMC, 0) \quad (351-43)$$

$$\sigma(RCMN) = \sqrt{(\max(\sigma(NHSO), 0))^2 + (\max(\sigma(NHNO), 0))^2 + (\max(\sigma(SOIL), 0))^2 + [1.8 * \max(\sigma(Cl^-), 0)]^2 + (\max(\sigma(LAC), 0))^2 + (\max(\sigma(OMC), 0))^2} \quad (351-44)$$

$$mdl(RCMN) = 0 \quad (351-45)$$

## 6. DATA VALIDATION

Data validation performed at UCD involves assessing the quality, reliability, and integrity of the data. Watson et al. (1995) define a three-level data validation process for environmental measurement studies. The levels are only intended as general guidelines. The IMPROVE data delivered to CIRA and AQS databases are considered to be a mixture of Level 1B and Level 2 validated data. The levels are applied to IMPROVE as follows:

**Level 0:** Data at this level are, in essence, raw data, obtained directly from the data acquiring instruments. These data can be reduced or reformatted, but are unedited and unreviewed, 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. Averaging times represent the minimum intervals recorded, and these data may need to be aggregated to obtain averages for the sampling periods. Level 0 data have not been edited for instrument downtime, nor have procedural adjustments for baseline shifts, span changes, or known problems been applied. IMPROVE Level 0 data includes:

- Raw pressure transducer and temperature data from the sampler flashcards or the V4 controllers before automated validity tests.
- Filter weight measurements before automated validity tests.
- XRF raw spectra.

**Level 1A:** Data at this level have passed several qualitative reviews for accuracy and completeness. The focus of Level 1A validation is to obtain as complete a data set as possible. IMPROVE Level 1A data validation includes:

- Reviewing operator log sheets to verify operation of the sampler.
- Verifying operator log sheet entries against sampler flashcard data.
- Assigning correct flow and temperature source codes.
- Assigning status flags to invalid or questionable samples to reflect sampler malfunctions, site or laboratory operator errors, or power outages.
- Identifying, investigating, and flagging data that are beyond reasonable bounds or that are unrepresentative of the variable being measured (e.g., flow rate measurements that change significantly over the sampling period).

**Level 1B:** Data at this level have passed additional automated quantitative and qualitative reviews for accuracy and internal consistency. Discrepancies that cannot be resolved are reported to the measurement laboratories for investigation. Data that deviate from consistency objectives are individually examined for errors. Obvious outliers (e.g., -85 °C temperature) are invalidated by applying a status flag. Changes to the data (e.g., swapping dates on consecutive samples) are recorded and documented by applying status flags and providing comments. Level 1B data review is carried out using custom software developed for this purpose. IMPROVE level 1B data validation includes:

- Verifying filter weight measurements to ensure that
  - the range is within specified limits;
  - the post-weight is greater than the pre-weight.
- Examining daily flow rates based on a report that identifies flow rates with significant variations over 24 hours.
- Setting status flags when deviations from nominal operational settings have occurred (e.g., flow rates outside quantitative tolerances).
- Examining the ion and carbon field blank data for evidence of sample swaps.
- Examining individual data points identified as potential sample swaps between two adjacent dates.
- Comparing the analytical data to expectations based on historical data.

**Level 2:** Level 2 data validation takes place after data from various measurement methods have been assembled in the UCD IMPROVE database. Level 2 validation involves cross-module comparisons of various species. Data submitted to CIRA and AQS databases are considered to be validated at Level 1B and Level 2. Additional Level 2 data validation is performed at CIRA.

IMPROVE Level 2 data validation consists of site-by-site and network-wide examination of time series and scatter plot of data, including:

- Comparing sulfur and sulfate concentrations.
- Comparing elemental carbon and light absorption coefficients.
- Examining PM<sub>10</sub> mass and PM<sub>2.5</sub> mass for cases where PM<sub>2.5</sub> is greater than PM<sub>10</sub>.
- Comparing PM<sub>2.5</sub> gravimetric mass and reconstructed mass.

**Level 3:** This level of data review is applied after data delivery and is beyond the scope of data validation performed by UCD. At this level, the data are reconciled with other research findings, such as modeling results or theoretical predictions. Level 3 validation continues for as long as the CIRA and AQS databases are maintained.

Data validation is not a linear process, and a significant amount of data validation is performed by the analytical laboratories before the data are delivered to the data validation specialist. The SOPs for the analytical laboratories describe their data validation procedures in detail. The following sections discuss the Level 1 and Level 2 validation processes that occur once the data are received from the field and laboratories.

## 6.1 DEFINITION OF STATUS FLAGS

Status flags are used as standardized abbreviations describing the status of individual sample results, and are assigned during the Level 1 and 2 validation processes (Table 11). Samples

associated with “Terminal” flag are invalidated for a variety of reasons, and no concentration values are reported, whereas those associated with “Informational” flag are still valid samples and concentrations are reported. The “Temporary” flags are replaced before final data reporting.

**Table 11.** Status flags and their definitions.

Status Flag	Description	Flag Type	AQS code
BI	Bad Installation of Sample Cartridge or Filter	Terminal	BJ
CG	Sample Flow Rate Out of Spec.	Informational	W
CL	Sample Flow Rate Out of Limits	Terminal	AH
DA	Sample not analyzed	Terminal	AM
DE	Reported value is an estimate	Informational	LJ
EP	Equipment Problem	Terminal	AN
LF	Sample Flow Rate Out of Spec.	Informational	W
NM	Normal	Informational	
NS	No Sample Collected/Late Sample Change	Terminal	AF
OL	Site Off Line	Terminal	AD
PO	Power Outage	Terminal	AV
QD	Questionable Data	Temporary	4
SA	Sampling Anomaly	Informational	1
SO	Still out	Temporary	
SP	Same-day Field Blank/Sample Swap	Informational	
SW	Sampling Dates Swap	Informational	
TU	Incorrect Time (with time shift >= 6hrs)	Informational	3
UN	Undetermined Weight	Informational	AM
XX	Sample Destroyed, Damaged or Contaminated	Terminal	AJ
PM	Undefined but allowed by SWAP as informational	No longer used	
NR	Not Reanalyzed by DRI	No longer used	
NA	Not Applicable	No longer used	AM
QA	Quality Assurance	No longer used	4
QC	Quality Control	No longer used	
RF	Really High Flow Rate	No longer used	W

## 6.2 LEVEL 1 VALIDATION PROCEDURES

Level 1 validation is conducted throughout the sample handling and analysis processes. Validation for the gravimetric PM<sub>2.5</sub> and PM<sub>10</sub> masses, PM<sub>2.5</sub> elements, ions, and carbon data is conducted by the laboratory technicians performing the analyses. The following Technical Information (TI) documents are available for mass validation and XRF data validation:

Mass validation: *Sample Handling TI 2510 General Laboratory Procedures*, Section 5.9

XRF validation: *XRF TI 130e Level I Validation*

Level 1 flow rate validation is performed as a four-step process. Additional Level 1B validation checks are performed on data completeness and field blank validity prior to processing the concentration data. The following sections discuss the flow validation and Level 1B checks in detail.

### 6.2.1 Flow Validation

Level 0 flow data (raw transducer readings from field log sheets and flashcards) are uploaded to the UCD IMPROVE database by the Sample Handling Laboratory technicians soon after sampled filters return to the lab from the IMPROVE sites. Prior to data uploading, as part of the Level 1A validation process, data from log sheets and flashcards are reviewed in parallel and checked for inconsistency resulting from sampling anomaly and/or sampler malfunction. In these cases, the sample status is changed from NM to a terminal or temporary flag, and filter/sample event comments are provided. When flashcard data is missing or not reliable due to equipment problems, the Flow Source Code (FlwSrc) for the affected sample is changed from default (MC) to log sheet (LC/LO) or nominal value (NF) to ensure accurate calculation of the average flow rate. Detailed procedures on flow data ingestion and Level 1A validation can be found in the *Sample Handling TI 251E Entering Log Sheets and Simple Problem Diagnosis*.

Several Level 1B checks on the 15-minute raw flow data are performed by running the *flow.check* function (for V2 controller data) and *flowV4.check* function (for V4 controller data) in the *datvalIMPROVE* R package. The functions return a report that lists the samples with abnormal flow variability, abnormal sampling temperature, and number of records for further investigation. The criterion for abnormal flow variability is greater than 8% during a 24-hour sampling period, which is due either to equipment/installation problems or to a steady pressure drop from heavy filter loading. Additional criteria are applied to the V4 controller data. The 15-minute raw pressure readings that are out of range ( $\text{Cyc\_pressure} < -1.25$  or  $> 1.25$ ;  $\text{Ori\_pressure} < 0$  or  $> 15$ ) are registered as NULL and excluded from the 24-hour average flow calculation. The 15-minute raw cyclone pressure readings that are slightly below 0 ( $-1.25 \leq \text{Cyc\_pressure} \leq 0$ ) are treated as 0 in the 24-hour average flow calculation.

Flow plots are carefully reviewed (IMPROVE Flow Graphs; <http://analysis.crocker.ucdavis.edu:3838/FlowRates/>) to determine flow validity. Proper Flow Source Code is assigned if the primary channel (MC) is not reliable.

Next, the 15-minute flashcard data are processed in SQL environment to derive the daily average flow rate and elapsed time (ET). The flow processing code automatically assign non-normal flow status flags to the samples that have flow rate that deviates from the nominal value or ET shorter than 18 hours. Table 12 and 13 list the types of flow flags and the associated criteria for applying them to PM<sub>2.5</sub> and PM<sub>10</sub> samples, respectively.

**Table 12.** Definitions and application criteria of automatic flow flags for PM<sub>2.5</sub>.

Automatic Flow Flag	Definition	Type	Criteria for Application for PM <sub>2.5</sub> Samples
CL	Clogged Filter	Terminal	Flow rate < 15 L/min for more than 6 hours if flashcard data are used Average flow rate < 15 L/min if log sheet values are used
CG	Clogging Filter	Informational	Flow rate < 18 L/min for more than 6 hours if flashcard data used Average flow rate < 18 L/min if log sheet values are used
LF	Low/high flow rate	Informational	Average flow rate < 19.7 L/min or > 24.1 L/min
PO	Power Outage	Terminal	Elapsed time < 1080 minutes (18 hours)

The cyclone characterization test carried out in the summer of 2016 on the IMPROVE PM<sub>2.5</sub> cyclones shows consistent results with the characterization performed by John and Reischl (1980). The particle size cut of the cyclone at any operating flow rate can be determined from the following equation:

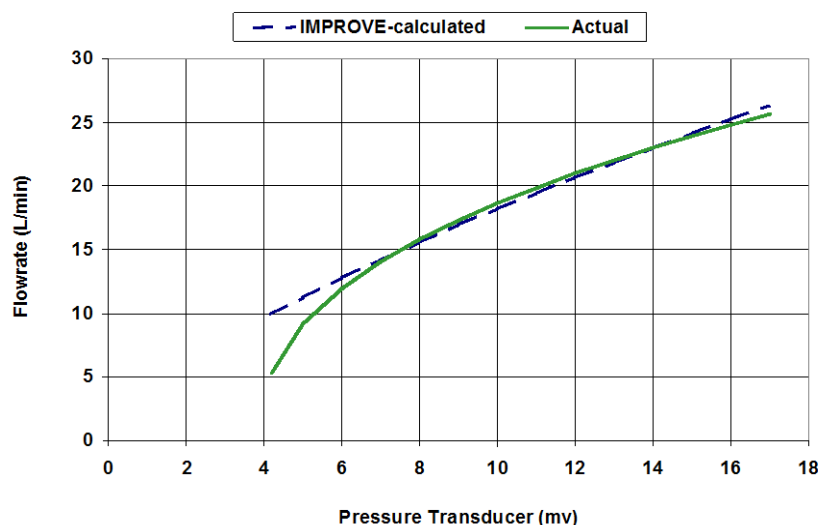
$$D_{50} = 52.5 * Q^{-0.99} \quad (351-46)$$

$D_{50}$  is the 50% cutoff diameter (in  $\mu\text{m}$ ) and  $Q$  is the flow rate (in L/min). Note that at the nominal flow rate of 23 L/min, the 50% cutoff diameter is 2.36  $\mu\text{m}$  rather than 2.5  $\mu\text{m}$ .

The criteria for the CL, CG, and LF flags are determined based on calculation limitations, performance testing, and particle size cut. Figure 4 illustrates a typical relationship between PM<sub>2.5</sub> flow rate and the cyclone pressure transducer measurement. The dashed line shows the calculated flow rate from equation (351-1) whereas the solid line shows the measured flow rate. The response of the actual flow rate to the change in pressure is no longer linear below approximately 15 L/min and therefore the calculated flow rate < 15 L/min is inaccurate. If >24 15-minute (6 hours in total) flow rate readings are below 15 L/min, or if the average flow rate is below 15 L/min when log sheet data are used, the sample is flagged as CL and no concentration data are reported. The PM<sub>2.5</sub> cyclone cut point is 3.6  $\mu\text{m}$  at 15 L/min.

The criteria for applying CG and LF flags are based primarily on cut point characterization of the PM<sub>2.5</sub> cyclone. The cut point is 3.0  $\mu\text{m}$ , 2.75  $\mu\text{m}$ , and 2.25  $\mu\text{m}$  at 18 L/min, 19.7 L/min, and 24.1 L/min, respectively. The 2.25 - 2.75  $\mu\text{m}$  range is considered a reasonable range of particle cut points for a data value to be labeled as PM<sub>2.5</sub>.

**Figure 4.** PM<sub>2.5</sub> sampler flow rate versus the cyclone pressure transducer reading for the IMPROVE-calculated flow rate and the measured (actual) flow rate.



A similar set of flags is applied to the PM<sub>10</sub> data (Table 13), but with several differences in the criteria, due principally to the lower flow rate at which the PM<sub>10</sub> sampler operates. The relationship between the PM<sub>10</sub> Sierra cyclone and particle size cut is not well characterized so the criteria are determined somewhat arbitrarily. It is important to note that under circumstance of a failing pump that produces less vacuum, equation (351-2) is no longer true and the calculated flow rates for the PM<sub>10</sub> module are not valid.

**Table 13.** Definitions and application criteria of automatic flow flags for PM<sub>10</sub>.

Validation Flag	Definition	Type	Criteria for Application for PM <sub>10</sub> Samples
CL	Clogged Filter	Terminal	Flow rate < 10 L/min for more than 6 hours if flashcard data are used Average flow rate < 10 L/min if log sheet values are used
CG	Clogging Filter	Informational	Flow rate < 14 L/min for more than 6 hours if flashcard data are used; Average flow rate < 14 L/min if log sheet values are used
LF	Low/high flow rate	Informational	Average flow rate < 15 L/min or > 18 L/min
PO	Power Outage	Terminal	Elapsed time < 1080 minutes (18 hours)

Finally, all samples flagged as terminal (i.e. CL and PO) by the flow processing code are manually reviewed for errors. In cases where valid samples are flagged as invalid (e.g. corrupt flashcard files or faulty transducer readings), the flow source code is changed and average flow rate reprocessed to correct the sample status.

### 6.2.2 Level 1B Checks

The analysis data reported by the measurement laboratories are ingested into the UCD IMPROVE database in their corresponding tables (i.e. analysis.Carbon, analysis.CarbonLaser, analysis.HIPS, analysis.Ions, and analysis.Mass), as described in Section 4. Several checks are performed using the *datvalIMPROVE* package in R, including:

- Data Completeness: the *completeness.check* function returns records with missing analytical data for each module.
- Field Blank Swap: the *ions\_fb.check* and *carbon\_fb.check* functions check for possible swap between same-day field blanks and samples for nylon and quartz filter samples.

Following the checks, concentrations, MDLs, and uncertainties are processed and posted in the analysis.Results table using the *improve\_calculate\_all* function in the *crocker* package. Additional checks on elapsed time and sampling days are performed by running the *etime.check* and *day.count* functions in *datvalIMPROVE*. These checks ensure there are no records with ET greater than 24 hours and no sites with less than 10 or more than 11 sampling days (February is typically an exception).

## 6.3 LEVEL 2 VALIDATION PROCEDURES

Level 2 validation is performed by comparing site-by-site concentration data obtained from different modules as well as by assessing network-wide long-term trends using a variety of R scripts and data visualization tools.

### 6.3.1 Cross-Module Comparison

#### 1A Module versus 2B Module

Quality assurance for the 1A and 2B Modules consists of comparing the measured concentrations of sulfur and sulfate. Sulfur concentrations are reported through elemental analysis of the PTFE filter from the 1A Module, while sulfate concentrations are determined by ion chromatography analysis of the nylon filter from the 2B Module. Discrepancies between 1A Module sulfur (times three, S3) and 2B Module sulfate (SO4) concentrations are investigated. If analytical error is suspected, a request is sent to the corresponding laboratories for a reanalysis of the sample.

The *swap.check* function in the *datvalIMPROVE* package returns samples marked as “swap” and/or “outlier”. For checking possible sample swaps, successive pairs of data are examined using the algorithm outlined below. In equation (351-47), two indices for each pair of sulfur and sulfate data are calculated using data from the current and the next sampling days (referred to as subscript 1 and 2, respectively).

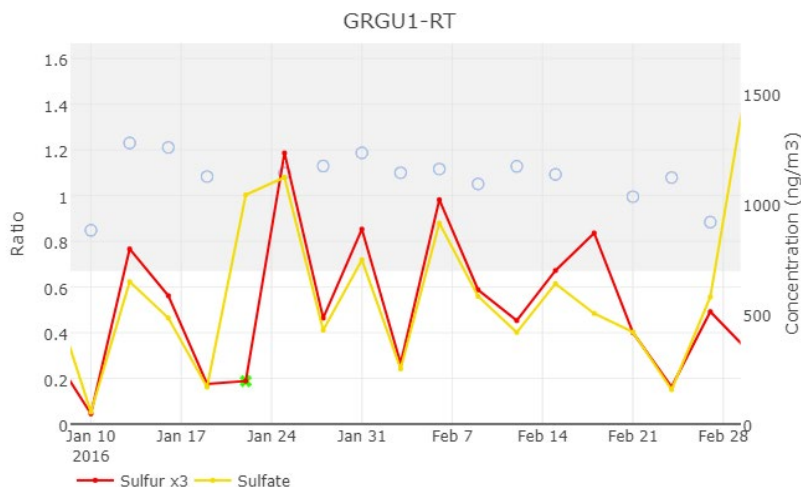
$$Index1 = \left( \frac{S3_1}{SO4_1} - 1 \right) \times \left( \frac{S3_2}{SO4_2} - 1 \right) \quad Index2 = \left( \frac{S3_1}{SO4_2} - 1 \right) \times \left( \frac{S3_2}{SO4_1} - 1 \right) \quad (351-47)$$

If PM<sub>2.5</sub> sulfur are in the form of sulfate, the S3/SO4 ratio is close to unity. If the samples are not subject to a swap, *Index1* would be close to zero and *Index2* would be large (and

may be either positive or negative). The criterion for flagging a pair of samples as swap is when  $Index1 < -0.03$  and  $-0.05 < Index2 < 0.05$ , which have been set empirically. The criterion for the “outlier” flag is when the S3/SO4 ratio  $< 0.667$  or  $> 1.8$ .

The S3/SO4 plot on the IMPROVE Data Site is used to further investigate samples flagged as swap and/or outlier. Figure 5 shows an example of an outlier pair at the GRGU1 site on 1/27/2016. On that day, the sulfate concentration is 1041.06 ng/m<sup>3</sup> while the S3 is 195.51 ng/m<sup>3</sup>, yielding a S3/SO4 ratio of 0.19, well below the acceptable range. In cases like this, the flow rate and elapsed time are first examined to make sure the correct flow source code is assigned. If analytical error is suspected, the XRF and/or IC laboratories perform a reanalysis. If the reanalysis results resolve the issue, the sample mass loadings are updated in the UCD IMPROVE database and the concentration data reprocessed. If the reanalysis results are the same as the original analysis, the samples may be flagged as terminal with XX (Sample Destroyed, Damaged or Contaminated) status.

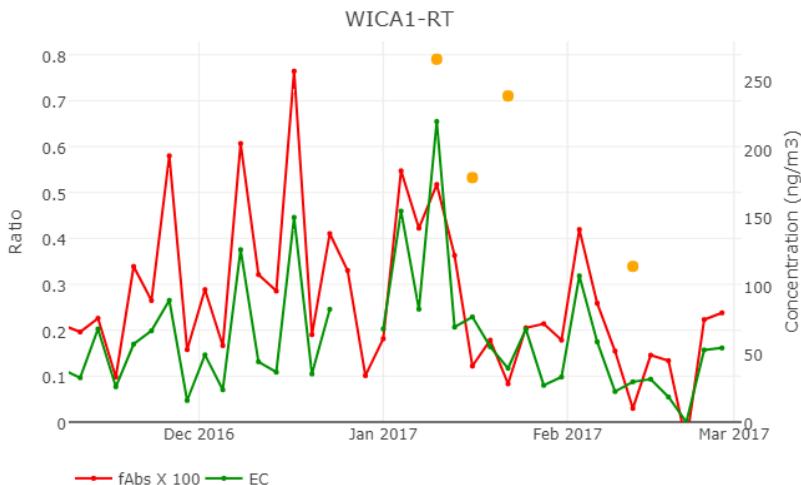
**Figure 5.** S3/SO4 comparison plot for the GRGU1 site showing the 1/21/2016 sample pair as an outlier (green x).



### 1A Module versus 3C Module

The light absorption coefficient ( $f_{abs}$ ) at 635 nm is measured by HIPS from the 1A Module PTFE filter and is compared qualitatively with the elemental carbon (EC) concentration measured by TOR from the 3C Module quartz filter. Visual inspection on the data is performed to identify outliers using the  $f_{abs}/EC$  plot on the IMPROVE Data site. The  $f_{abs}$  and EC are typically correlated to some extent (Figure 6). If analytical error in either measurement is suspected, other measurement data from the same module would be examined together to determine validity of the sample.

**Figure 6.** Comparison plot of light absorption coefficient measurements ( $f_{abs}$ , times 100) from 1A Module and elemental carbon (EC) measurements from 3C Module at WICA1 site.



### 1A Module versus 4D Module

1A module  $PM_{2.5}$  mass and 4D module  $PM_{10}$  mass are reviewed and compared (Figure 7). The *mf\_mt.check* function in the *datvalIMPROVE* package returns a list of samples flagged as mass outliers if any of the following criteria are met:

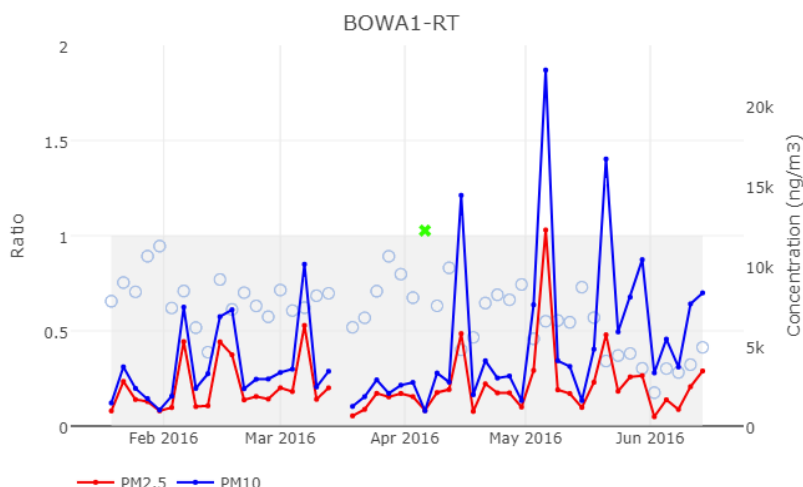
- $PM_{2.5}$  or  $PM_{10}$  mass concentration is negative.
- $PM_{2.5}$  mass is greater than  $PM_{10}$  mass and  $Z\_score > 1$ .
- $PM_{10}$  mass is abnormally high and  $Z\_score > -43$  (the number 43 is set empirically).

$Z\_score$  is calculated using equation (351-48)

$$Z\_score = 1.41 \times \frac{PM_{2.5} - PM_{10}}{\sqrt{(unc_{PM_{2.5}})^2 + (unc_{PM_{10}})^2}} \quad (351-48)$$

For samples that are flagged for one of the above cases, the filter is reweighed to confirm the post-weight; the pre-weight cannot be re-determined after sampling. Samples with invalid mass concentrations are flagged as “UN” (Undetermined Weight).

**Figure 7.** Time series plot of PM<sub>10</sub> and PM<sub>2.5</sub> masses and their ratio at BOWA1 site.



### PM<sub>2.5</sub> reconstructed mass versus gravimetric mass

The PM<sub>2.5</sub> reconstructed masses, RCMC and RCMN, are calculated by equations 351-40 and 351-43, respectively. RCMC and RCMN are compared to the gravimetric mass (MF) as a check of all measured components from the 1A, 2B, and 3C Modules (Figure 8). The *mf\_rmc.check* function in the *datvalIMPROVE* package returns a list of samples flagged as outliers if any of the following criteria are met:

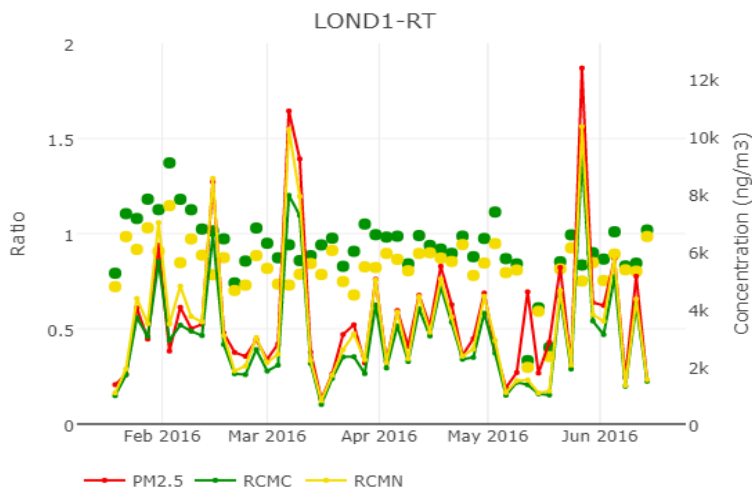
- RCMC is higher than two times MF, and the RCMC Z<sub>score</sub> > 3; the number 3 is set empirically. These samples are accompanied with a comment “MF << RCMC”.
- The RCMN Z<sub>score</sub> < -22; the number 22 is set empirically. These samples are accompanied with a comment “MF >> RCMN”.

Z scores are calculated as follows:

$$RCMC\_Z\_score = 1.41 \times \frac{RCMC - PM_{2.5}}{\sqrt{(unc_{PM_{2.5}})^2 + (unc_{RCMC})^2}} \quad (351-49)$$

$$RCMN\_Z\_score = 1.41 \times \frac{RCMN - PM_{2.5}}{\sqrt{(unc_{PM_{2.5}})^2 + (unc_{RCMN})^2}} \quad (351-50)$$

**Figure 8.** Time series plot of PM<sub>2.5</sub> gravimetric mass, reconstructed mass without nitrate (RCMC) and reconstructed mass with nitrate (RCMN) and their ratios at LOND1 site.

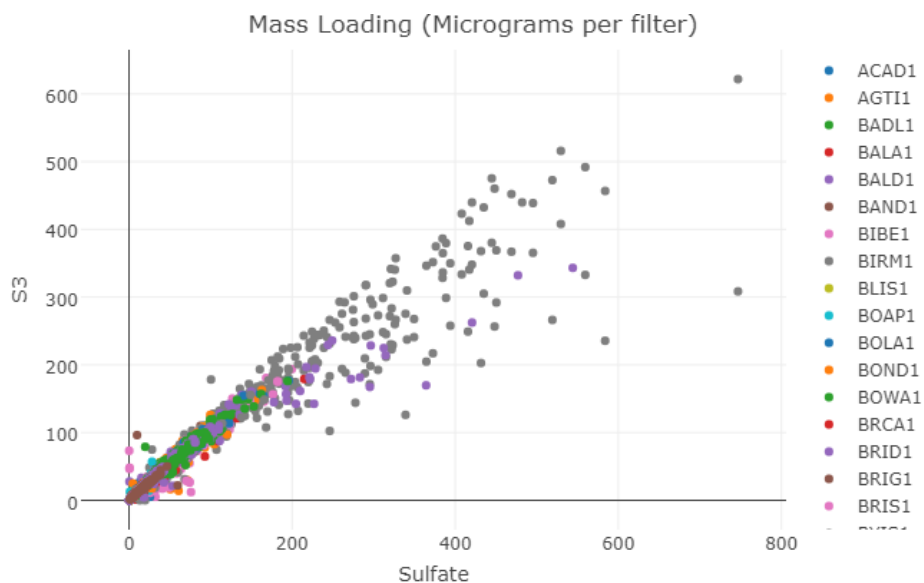


### 6.3.2 Long-Term Network-Wide Checks

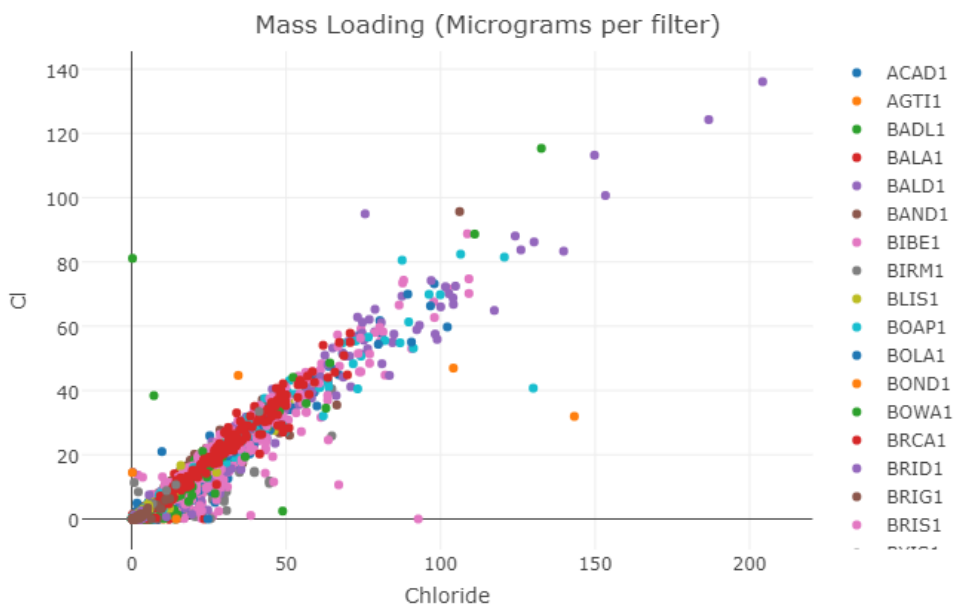
Several data visualization tools and control plots are used for long-term network-wide checks in addition to the site-by-site monthly data evaluation. These checks help reveal the long-term trends and seasonal patterns if any, as well as any network-wide problems. Below are examples of the tools and plots that are routinely used and reviewed:

- Scatter plot of S3 versus SO<sub>4</sub> mass loadings for the whole network (Figure 9). This plot is accessible from the IMPROVE Data site, "Early Review" tab.
- Scatter plot of chlorine versus chloride mass loadings for the whole network (Figure 10). This plot is accessible from the IMPROVE Data site, "Early Review" tab.
- Scatter plot of  $f_{\text{abs}}$  versus BC (converted from TOR absorption measurements) for the whole network (Figure 11). This plot is accessible from the IMPROVE Data site, "Early Review" tab.
- Time series plot of the 1A to 4D mass loading ratio showing the long-term trend and historical data at a given site (Figure 12). This tool is accessible from the IMPROVE Data site, "Mass Review" tab.- Monthly median, 90%, and 10% percentiles of the concentration data for all reported species. Figure 13 shows an example time-series plot for OC concentrations between 2011 and 2016. These plots are generated in R.

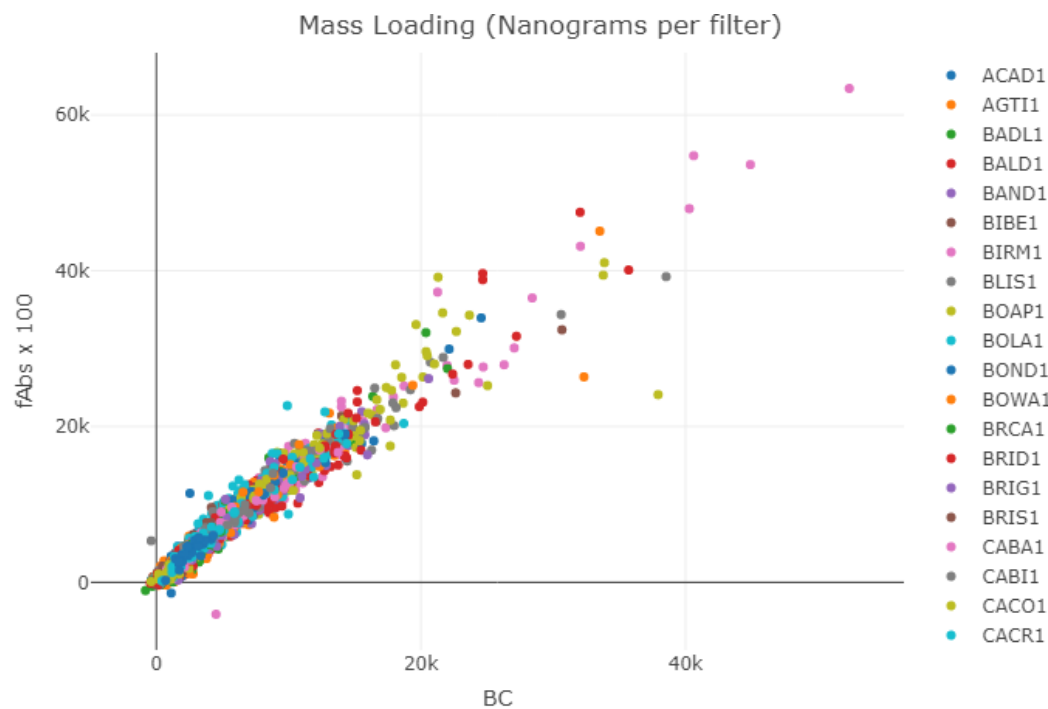
**Figure 9.** Scatter plot of sulfur ( $\times 3$ ) versus sulfate for the entire IMPROVE network.



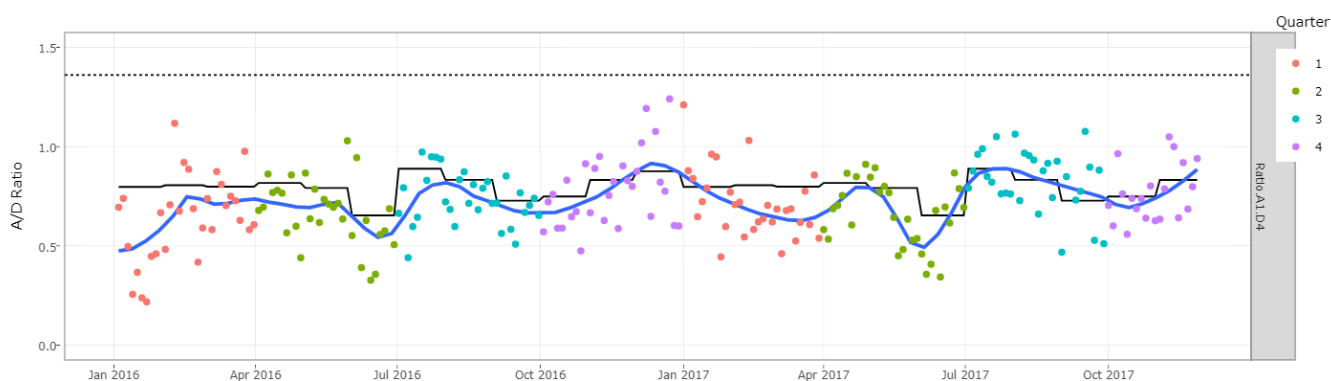
**Figure 10.** Scatter plot of chlorine versus chloride for the entire IMPROVE network.



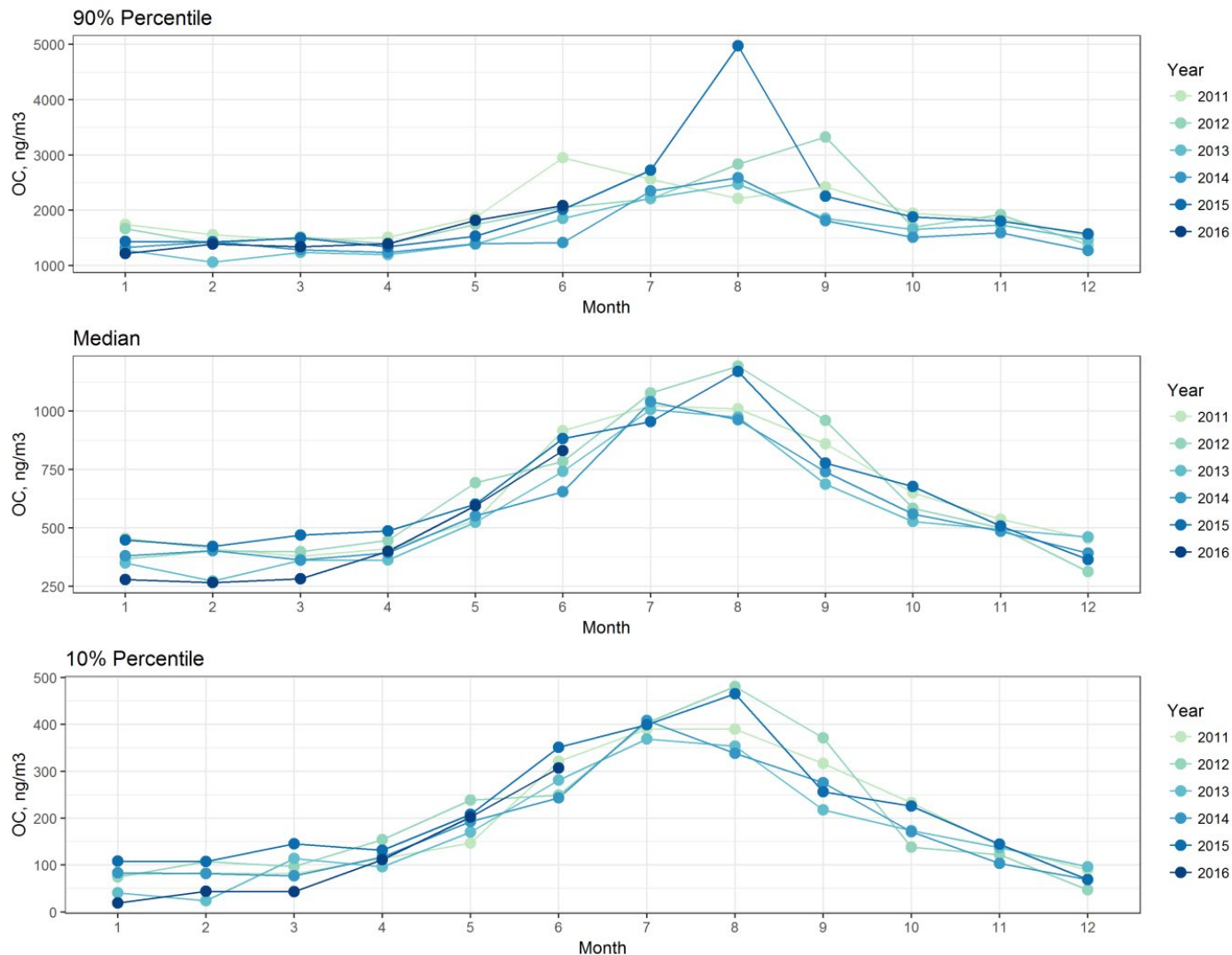
**Figure 11.** Scatter plot of  $f_{abs}$  versus BC (converted from TOR absorption measurement) for the whole network.



**Figure 12.** Ratio of PM<sub>2.5</sub> mass (1A) over PM<sub>10</sub> mass (4D) at ACAD1 site, represented as raw measurements not adjusted for flow rates. Points are individual sample days (pink = Q1, green = Q2, blue = Q3, purple = Q4). Black line is the multi-year monthly mean. Blue line is the locally weighted average (LOESS).



**Figure 13.** Multi-year monthly 10% percentile (top), median (middle) and 90% percentile (bottom) of organic carbon (OC) concentrations (in ng/m<sup>3</sup>) for the whole IMPROVE network from 2011 to 2016.



### 6.3.3 Final Review

Several final checks are performed before submission of data delivery files to the CIRA, AQS, and UCD CIA databases:

- The *QD.check* function in *datvalIMPROVE* returns a list of the samples flagged with QD for a specified month. The QD status is normally assigned by the sample handling lab technicians during initial inspection of the physical samples and the raw flow rate data. These cases are investigated and resolved by either changing the status back to NM or replacing the QD with appropriate terminal or informational flags. There should be no records with QD in the status field in the delivery files.

- The *ObjCode.check* function in *datvalIMPROVE* performs a check on the ObjectiveCode field in the data file. This field should only contain RT (routine) or CL (collocated). The *ValidSta\_NullData* function in *datvalIMPROVE* checks if there are valid data (instead of -999) for any valid samples.
- The *Validsta\_BadData* function in *datvalIMPROVE* uses a set of criteria listed in the R code to check if any valid sample has concentration data that are out of normal range and returns a list of samples for further investigation.
- The *MDL\_UNC* function in *datvalIMPROVE* checks if any calculated MDLs or uncertainties have negative values.
- The *final.check* function in *datvalIMPROVE* returns a list of samples that are either invalid with valid concentrations (not -999) or valid with -999 as concentrations in a specific FED delivery file.
- The *sitecount* function in *datvalIMPROVE* returns the site count for a specific FED delivery file.

## 7. DATA DELIVERY

After Level 2 data validation is complete, the data files are submitted to CIRA, AQS, and UCD CIA databases.

### 7.1 SUBMISSION TO CIRA

FED export files are created using the *improve\_export\_fed* and *improve\_export\_wide* functions in the *crocker* package, in which the year, month, and server for both functions are entered. The functions create “skinny” and “wide” versions of the dataset, and both are submitted. The files are saved under *U:\IMPROVE\FED Export*, named ‘IMPROVE\_Data\_YYYY\_MM\_server’ and ‘IMPROVE\_WideData\_YYYY\_MM\_server’ (e.g., “IMPROVE\_Data\_2017\_02\_production”), respectively. These files are compressed into a zip folder and are emailed to the CIRA correspondent(s) as an attachment.

### 7.2 SUBMISSION TO AQS

Data files are prepared and delivered to AQS following these steps:

1. Create the AQS delivery files using the IMPROVE Management Site, ‘Analysis Data’ tab. Choose **AQS** for the ‘Format’ and fill in the Year, Start Month and End Month.
2. Click continue to automatically generate the file. Save the file to the UCD U Drive: *U:\IMPROVE\AQS\AQS Export*.
3. Open a web browser and navigate to the EPA Exchange Network Services website, <https://enservices.epa.gov/login.aspx>. Use credentials to login.
4. From the home screen, search for the AQS submission form by clicking on the **Go** button of the Exchange Network Services bar. Type “AQS” into the search bar. The search results will show all available processes associated with the AQS system. Choose the

service that has “AQS Submit” specified in the “Service Name” field. This will take the analyst to the AQS submission form.

5. Fill out the submission form, specifying email address, AQS user ID, screening group (IMPROVE), the file type (FLAT), the final processing step (POST), and whether or not to stop on errors (YES). Use the **Choose File** button to select the file generated from the previous step. Press the **SEND DATA** button to submit the form.
6. Monitor progress of the data submission through the same web portal.

### 7.3 SUBMISSION TO UCD CIA

1. Open a web browser and navigate to the UCD CIA submission website, <https://ci1A.uploadportal.azurewebsites.net/>. Use credentials to login.
2. Click the ‘Continue’ button in the center of the page.
3. Specify the network of choice that you will be delivering the data for, which in this case we be ‘IMPROVE’.
4. Click ‘Browse’ and select the file generated/submitted successfully to AQS.
5. Once the file is selected, click ‘Submit’; the next page will indicate if the submission was successful.

## 8. REFERENCES

- Hyslop, N.P. and White, W.H. (2008) Estimating Precision Using Duplicate Measurements. J. Air & Waste Manage. Assoc. 59:1032–1039.DOI:10.3155/1047-3289.59.9.1032.
- John, W. and Reischl, G.P. (1980) A Cyclone for Size-Selective Sampling of Ambient Air, J. Air Pollut. Control Assoc., 30 (8), 872-876.
- Watson, J.G.; Liroy, P.J.; Mueller, P.K. (1995). The measurement process: Precision, accuracy, and validity. In Air Sampling Instruments for Evaluation of Atmospheric Contaminants, 8th; Cohen, B. S., Hering, S. V., Eds.; American Conference of Governmental Industrial Hygienists: Cincinnati, OH, 187-194.