UCD IMPROVE Technical Information #351E

Flow Validation

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

October 4, 2022
Version 1.0

Prepared By: DocuSigned by: Indu Thekkemeppilly Sivakumar 19F6B63B1B17443... Date: 10/7/2022

Reviewed By: DocuSigned by: Dominique Young BB55DBA34BAB407... Date: 10/7/2022

Approved By: DocuSigned by: Marcus Langston 0A10CECE79B0452... Date: 10/7/2022

UCDAVIS
AIR QUALITY RESEARCH CENTER

DOCUMENT HISTORY

Date Modified	Initials	Section/s Modified	Brief Description of Modifications
03/14/22	SRS	All	Previously anthologized version separated into individual TIs; content updated.

TABLE OF CONTENTS

LIST OF TABLES

1. PURPOSE AND APPLICABILITY

The purpose of this technical information (TI) is to provide information regarding the steps to process and validate the flow data from the Interagency Monitoring of Protected Visual Environments (IMPROVE) network. Flow data from the network are reviewed and validated using various tools.

2. SUMMARY OF THE METHOD

The University of California, Davis (UCD) analyst uses the UCD Flow Plotter website along with custom software in the R language to perform flow data processing and validation.

3. DEFINITIONS

- **AQRC:** Air Quality Research Center.
- **crocker:** A custom software package in the R language that contains the data processing code used to produce, check, and post the final results.
- **datvalIMPROVE:** A custom software package in the R language that contains the data validation code used to collect, compare, and flag the final results.
- **Energy Dispersive X-Ray Fluorescence (EDXRF):** An analytical technique used to determine the concentration of elements.
- **Hybrid Integrating Plate/Sphere (HIPS):** An analytical technique for optical absorption.
- **Interagency Monitoring of Protected Visual Environments (IMPROVE):** Federal $PM_{2.5}$ and PM_{10} sampling network directed by the National Park Service, with sites located principally in remote rural areas.
- **IMPROVE database:** A SQL Server database that is the central warehouse of IMPROVE preliminary and final data at UCD.
- **NPS:** National Park Service.
- **PM:** Particulate Matter. $PM_{2.5}$ is particulate matter with diameters 2.5 micrometers (μm) and smaller. PM_{10} is particulate matter with diameters 10 μm or smaller.
- **SOP:** Standard Operating Procedure.
- **SQL:** database management system used by AQRC.
- **TI:** Technical Information; subset document paired to an SOP.
- **UCD:** University of CA—Davis.

4. HEALTH AND SAFETY WARNINGS

Not applicable.

5. CAUTIONS

Not applicable.

6. INTERFERENCES

Not applicable.

7. PERSONNEL QUALIFICATIONS

The UCD Air Quality Research Center (AQRC) Data and Reporting Group staff assigned to tasks described in this document have advanced training in database programming and database management.

8. EQUIPMENT AND SUPPLIES

The hardware and software used for IMPROVE data validation are described in the associated *UCD IMPROVE SOP #351: Data Processing and Validation*.

9. PROCEDURAL STEPS

Flow data from the V4 controllers is automatically transmitted daily to the UCD IMPROVE database for near real-time review by the Sample Handling Laboratory (SHL) and Field Group. Field log sheets and flashcards (with raw pressure transducer readings) are also available as backup flow data and are shipped with the physical sampled filters from the field sites to the UCD SHL. The SHL receives flow data from the V2 controllers by flashcard and log sheet; only one IMPROVE site has the older V2 controller (BYIS). As part of the Level 1A validation process, flow data are reviewed 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 automatically transmitted flow data are not available, the flashcard, log sheet, or nominal value can be used instead. The Flow Source Type Code for the affected sample is changed from the default (MC/MO) to log sheet (LC/LO) or nominal value (NF) to ensure an accurate calculation of the average flow rate. Detailed procedures on flow data ingestion and Level 1A validation can be found in *UC IMPROVE TI #251E: Entering Log Sheets and Simple Problem Diagnosis*.

9.1 Processed Flow Data

Prior to checking flow data, the quality assurance officer processes flow data using the SQL query or the *improve_process_flow* function in R as described in section 9 of *UCD IMPROVE TI #351B: Data Processing* to derive the daily average flow rate and elapsed time (ET). The flow processing code automatically assigns non-normal flow status flags to the samples with flow rates that deviate from the nominal values. Table 1 and 2 list the

types of flow flags and the associated criteria for applying them to $PM_{2.5}$ and PM_{10} samples, respectively.

Table 1. Definitions and application criteria of automatic flow flags for $PM_{2.5}$.

Automatic Flow Flag	**Definition**	**Type**	**Criteria for Application for $PM_{2.5}$ 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)
EP	Equipment Problem	Terminal	Elapsed time > 1800 minutes (30 hours) or is missing
TO	Timing Outside normal bounds	Informational	Elapsed time between 1080 minutes (18 hours) - 1380 minutes (23 hours) or 1500 minutes (25 hours) – 1800 minutes (30 hours)

The 2016 IMPROVE $PM_{2.5}$ cyclone characterization test yielded results consistent 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:

$$D_{50} = 52.5 * Q^{-0.99} \quad \text{(351E-1)}$$

Where,

D_{50} = 50% cutoff diameter (in µm)

Q = flow rate (in L/min)

Note that at the nominal flow rate of 23 L/min, the 50% cutoff diameter is 2.36 µm rather than 2.5 µm.

The criteria for the CL, CG, and LF flags are determined based on calculation limitations, performance testing, and particle size cut. 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 $PM_{2.5}$ cyclone cut point is 3.6 µm at 15 L/min.

The criteria for applying CG and LF flags are based primarily on cut point characterization of the $PM_{2.5}$ cyclone. The cut point is 3.0 µm, 2.75 µm, and 2.25 µm at 18 L/min, 19.7 L/min, and 24.1 L/min, respectively. The 2.25 - 2.75 µm range is considered a reasonable range of particle cut points for a data labeled as $PM_{2.5}$.

A similar set of flags is applied to the PM_{10} data (Table 2), but with several differences in the criteria, due principally to the lower flow rate at which the PM_{10} sampler operates. The relationship between the PM_{10} 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_{10} module are not valid.

Table 2. Definitions and application criteria of automatic flow flags for PM_{10}.

Validation Flag	**Definition**	**Type**	**Criteria for Application for PM_{10} 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)
EP	Equipment Problem	Terminal	Elapsed time > 1800 minutes (30 hours) or is missing
TO	Timing Outside normal bounds	Informational	Elapsed time between 1080 minutes (18 hours) - 1380 minutes (23 hours) or 1500 minutes (25 hours) – 1800 minutes (30 hours)

Once the flow data have been processed, the data are to be validated. The following sections describe the procedure for generating a report containing flow related items that have met check criteria and require further investigation as well as some commonly observed scenarios.

Several Level 1B checks (see *UCD IMPROVE TI #351C: Data Validation* for details on Level 1B) on the 15-minute raw flow data are performed by running the *flow.check* function (for both the V2 and V4 controller data) from the *datvalIMPROVE* R package. To perform these checks, open an R environment (such as RStudio) and run the following command:

[month_flow] <- datvalIMPROVE::flow.check(startdate = [‘YYYY-MM-DD’], enddate = [‘YYYY-MM-DD’], site = [‘%’], list_all = [‘FALSE’], server = ‘production’)

When *list_all* is set to FALSE, the function returns a report that lists all the samples during the date period specified with abnormal flow variability, abnormal sampling temperature, and number of records for further investigation. If the *list_all* argument is set to TRUE, only the sample events with relative standard deviation out of range will be returned. The three asterisks (***) are generated automatically in the output from the *flow.check* function to indicate data issues.

The analyst can perform the checks for all active sites in the network by setting *site = '%'* or just for a particular site by specifying the site name. Several criteria are checked:

- Abnormal flow variability: > 8% during a 24-hour sampling period; can be caused by equipment installation problems or steady pressure drop from heavily loaded filter.
- Abnormal sampling temperature: relative standard deviation of temperature < 0.01% or > 10%; average temperature < 20 °C or > 40 °C.
- Abnormal number of records: number of 15-minute flow readings is < 72 rows (equivalent to 18 hours of run time) or > 104 rows (equivalent to 26 hours of run time).

Additional criteria implemented for the V4 controller include:

- The 15-minute raw pressure readings that are out of range (CYC pressure < -1.25 or > 1.25; 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 ≤ CYC pressure ≤ 0) are treated as 0 in the 24-hour average flow calculation.

9.2 Generating the Flow Validation Report

The flow validation report is generated as an Excel spreadsheet. It is populated using the data returned from running several checks on the flow data. As the first step of validation, check for valid filters with missing flow data. The *flow.completeness* check will return a list of filters with missing flow data.

- No Flow data: To generate the list, run the following command in the R environment:

```
[No flowdata] <- datvalIMPROVE::flow.completeness(startdate =
['YYYY-MM-DD'], enddate = ['YYYY-MM-DD'], server = 'production')

write.csv(No flowdata, "U:/IMPROVE/Data_Validation/Flow/Nofllow.csv",
row.names = TRUE)
```

  Once the list is generated, coordinate with the Sample Handling Laboratory to investigate the reason(s) behind the missing flow data and resolve as appropriate. Once all the filters have the correct flow data attached, reprocess the flow using the SQL query or the *improve_process_flow* function in R as described in Section 9 of IMPROVE TI 351B.

The next tab of the spreadsheet is populated using the data returned from running the *flow.check* function as described in section 9.1 above. The spreadsheet has several tabs as described below:

- V2 Controller Flow Review: This sheet is populated with flow data from sites still using the V2 controller (e.g. BYIS1). Generate this data by running the following command in R:

  View([month_flow]$OldController)

  Save the data frame as a CSV file using the following R command:

  write.csv(month_flow$OldController, "U:/IMPROVE/Data_Validation/Flow/Monthflow_OldController.csv", row.names = TRUE)

Once all the flow validation-related data frames are exported (the steps are below) in CSV format, they can be combined to Excel format to make the flow validation report. Once in Excel format, color code the modules (A = red, B = Yellow, C = Green, and D = Blue). The three asterisks (***) generated automatically in the output from the *flow.check* function (see Section 9.1) indicate data issues.

- V4 Controller Flow Review: This sheet is populated using flow data from sites using the V4 controller. Generate this data by running the following command in R:

  View([month_flow]$NewController$MainCheck)

  As described in the previous step, export the data frame and color code the modules.

- V4 Controller Solenoid Check: This sheet is populated with flow source records for cases where the open solenoid position is not equal to the cartridge position. Generate this data by running the following command in R:

  View([month_flow]$NewController$SolenoidCheck)

- Flow flags (CG, CL, LF, PO, EP, TO): These sheets contain lists of samples where the flow status is flagged as CG, CL, LF, PO, EP, or TO and require confirmation of appropriate flagging (see Tables 1 and 2). Generate this data by running the following command in R:

  [month_flowflag] <- datvalIMPROVE::flow.status(startdate = ['YYYY-MM-DD'], enddate = ['YYYY-MM-DD'], flowflag = [('CG', 'CL', 'LF', 'PO', 'EP', 'TO')], server = 'production')

  To generate a list with only one of the flow flags, set the *flowflag* argument to equal one of the six flags. Export the results and add it to the appropriately labelled sheet in the flow validation report.

9.3 Flow Validation

To further investigate the data returned from the flow checks and to validate flow data, flow plots are carefully reviewed (IMPROVE Flow Graphs; https://shiny.aqrc.ucdavis.edu/FlowRates/). The Flow Source Code is assigned if the primary source (MC for A, B, C modules and MO for D module; automatically

transmitted flow data or flash card) is not reliable. Guidelines for validating flow data include:

- Review the flow graph to identify unstable flow readings. Evaluate to determine if there is an absence of pattern or if the flow is changing gradually during the sampling day. No pattern indicates a potential issue requiring further investigation. Gradual change throughout the sampling period may be caused by heavy loading.
- If automatically transmitted flow data and flashcard data are not available or reliable, use log sheet data which can be retrieved from [Improve_2.1].[ops].[ControllerFilterReadings] or the hand-written records on the paper Field Log Sheets.
- The Flow Source Code or Filter Status Code can be updated as needed from the Filters page of the IMPROVE Management Site.
- Utilize the Average Flow Plot in the Flow Graphing App to further evaluate flow data.
- Utilize the Early Review page in the IMPROVE Data App to view site-by-site analysis data, which can be used to help evaluate flow issues.
- Utilize the Controller Filter readings page (https://improve.aqrc.ucdavis.edu/Operations/ControllerFilterReadings) of the IMPROVE Management Site.

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 flash card files or faulty transducer readings), the flow source code is changed and average flow rate is reprocessed to correct the sample status. The same approach is taken for other flow flags like CG and LF.

9.3.1 Common Flow Review Scenarios

In this section, common scenarios investigated during flow validation are described including guidelines for resolving issues.

Clogged / Clogging (CL or CG) status

- The flow data are flagged with CL or CG status when there is heavy loading on the filter or due to pump malfunction. In the case of heavy filter loading, no further action is needed. If the flow status is CL due to pump malfunction, change the filter status code to EP (Equipment Problem. Refer Table 1 and Table 2) from the Filters page of the IMPROVE Management Site and reprocess the flow data using the SQL query described in section 9 of *UCD IMPROVE TI #351B: Data Processing*.

Power Outage (PO) status

- The PO status gets applied when Elapsed time < 1080 minutes. Check the filter readings table, flow plotter, or log sheet data to ensure a late sample change was

not the cause for the elapsed time to fall below the limit. If the sample change was late, update the filter status to NS (No Sample. Refer Table 1 and Table 2) and reprocess the flow data using the SQL query described in section 9 of TI #351B.

Temperature probe malfunction

- If the temperature data is showing extreme values (e.g., 200 degrees Celsius), the Temperature probe could be malfunctioning. If the flow data looks normal and analysis values look good, this can be confirmed as a malfunction. Check the temperature data from nearby sites and or local weather records available online to rule out extreme events. In such cases, we can use the nominal temperature for flow calculations. The temperature source code can be updated to Nominal from the Filters page of the IMPROVE Management Site and reprocess the flow data using the SQL query described in section 9 of TI #351B.

Low Flow (LF) flow status

- Common reasons for the LF status are:
    - Heavy loading which results in a flow value between NM range and CG range as described in Table 1 and Table 2. No action is required in such cases.
    - Swaps between filter types. If consecutive sample dates have an LF flag and the filters are all in the same cartridge, check if any other module is affected by flow fluctuation. If there is, it is suspected that the cartridges have been swapped between the modules. Examine the following dates to see if the flow pattern returned to normal after a sample change. In such cases, the filter statuses can be updated to QD (Questionable Data) from the Filters page of the IMPROVE Management Site for further review after all analyses come back. If the pattern is continuing, request the field group or sample handling lab to contact the site operator to ensure proper installation of cartridges.

Double filter

- Double filters is most commonly observed for 3C Module filters. If low flow (lower range as in Table 1 and Table 2) is isolated to a single day, the cause could be because a double filter was loaded. If the SHL has observed double filters or extra screens on the cassette during download, the QD (Questionable Data) status is applied to the filter. If there is no explanation for low flow, flag the filter with the QD status at this point. In both cases, the data require further review once all analysis results have been received and the data can be processed and validated.

10. DATA AND RECORDS MANAGEMENT

The IMPROVE data are stored in Microsoft SQL Server Databases at UC Davis. The production database is run on a dedicated Windows Server with a RAID array for storage

and with offsite backups. Our development and test database environments are virtual machines. To test back up recovery, our development and testing environments are regularly restored from the production backups.

Data management is handled through custom software that interfaces with the UCD IMPROVE database. The primary applications for data ingest and management were developed on the .NET platform. Data processing and calculations were developed as R software packages. In addition, to support data validation and operational monitoring, several interactive visualizations have been developed using the R Shiny platform.

10.1 Disaster Recovery Plan

The scope of recovery activities will depend on the nature of the disaster. Response to an actual disaster may require implementing multiple sections of this SOP.

10.1.1 Facility Recovery

Private security services patrol the laboratory building on a regular basis (including nights, weekends, and holidays). In addition, campus facilities and maintenance staff are on call at all times.

Databases, file servers, and web server virtual and dedicated machines operate primarily out of the Metro IT data center in Hoagland Hall on the UCD campus. Metro IT has a highly-available, disaster recoverable virtualization environment. Weekly backups of the virtual hard drives are taken offsite and stored in the Campus Data Center. In the event of a disaster in Hoagland, critical machines will be mounted at the Campus Data Center. The Drew Avenue laboratory is directly connected to the main campus internet. In the event that connection is disrupted (such as through a construction accident), connections will be switched to a local backup server until service can be restored.

10.1.2 Hardware Recovery Plan

The campus network of IT Administrator staff allows for rapid response to server failure and recovery issues.

10.1.3 Software and Data Recovery Plan

10.1.3.1 UCD Laboratories

Raw and processed analysis data produced with the UCD laboratories are saved and available for use at any time on the computers associated with each instrument, including the PANalytical Epsilon 5 EDXRF, MTL Automated Weighing System (gravimetric mass), Hybrid Integrating Plate and Sphere (HIPS).

Operational flow rate information from samplers in the field is automatically transferred nightly to a file processing server. As a backup, the flow data are stored on SD cards and delivered to the sample handling lab along with the exposed filters.

Data from all analyses, along with the flows, are scheduled to automatically transfer to a central Microsoft SQL Server database located at a data center on the UCD campus. Differential backups are performed daily, and full backups are performed weekly.

10.1.4 Data Security

UCD access policies: Access to databases and computers associated with this project is limited to authorized project personnel by use of access control lists for files, programs, and database access. Access to laboratory and office space is controlled by keycards.
Password policies: Unique passwords are issued to each employee by the UCD campus system administrator. Password integrity is monitored by the UCD campus system administrator.
Termination policies: System access is revoked for terminated personnel. The IT Administrator disables domain accounts and passwords upon termination of employment.
Virus protection: Microsoft Endpoint Protection is used for virus scanning and protection. All staff are required to complete annual cyber security awareness training.

11. QUALITY ASSURANCE AND QUALITY CONTROL

11.1 Code Development

Software for data management, processing, and validation is developed in-house by professional software engineers. Source code is managed through a code repository. Development of code changes and new applications is conducted on a development environment that parallels the production environment. Prior to deployment in production, all code changes undergo testing within a separate test environment. The testing, which is conducted by developers, managers, and users, is targeted both at the identification of software bugs and the confirmation of valid data equivalent to the production system.

11.2 Bug Reporting

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

11.3 Data Validation

Data integrity is enforced within the UCD IMPROVE database via unique primary keys and non-nullable records. Data completeness and data quality are thoroughly checked through the data validation process, as described elsewhere in this TI.

12. REFERENCES

Walter John & Georg Reischl (1980): A Cyclone for Size-Selective Sampling of Ambient Air, Journal of the Air Pollution Control Association, 30:8, 872-876, DOI: 10.1080/00022470.1980.10465122