DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

UCD IMPROVE Standard Operating Procedure #351

Data Processing and Validation

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

September 30, 2022
Version 6.0

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 0A10CFCF79B0452... Date: 10/7/2022

UCDAVIS
AIR QUALITY RESEARCH CENTER

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

DOCUMENT HISTORY

Date Modified	Initials	Section/s Modified	Brief Description of Modifications
04/24/21	SRS	All	Reformatted to fit document guidelines.
05/04/22	SRS	3, 8, 9	Updated terms, added content due to reorganization of SOP.

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

TABLE OF CONTENTS

1. Purpose and Applicability 4
2. Summary of the Method 4
3. Definitions 5
4. Health and Safety Warnings 6
5. Cautions 6
6. Interferences 6
7. Personnel Qualifications 6
   7.1 Data & Reporting Group Manager 7
   7.2 Lead Quality Assurance Officer 7
   7.3 Quality Assurance Officer 7
8. Equipment and Supplies 7
9. Procedural Steps 9
10. Data and Records Management 10
   10.1 Data Disaster Recovery Plan 10
      10.1.1 Facility Recovery 10
      10.1.2 Hardware Recovery Plan 10
      10.1.3 Software and Data Recovery Plan 11
      10.1.4 Data Security 11
11. Quality Assurance and Quality Control 11
   11.1 Code Development 11
   11.2 Bug Reporting 11
   11.3 Data Validation 12
12. References 12

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

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 Interagency Monitoring of Protected Visual Environments (IMPROVE) network. Data processing and data validation are performed in parallel.

This Standard Operating Procedure (SOP) broadly outlines the procedures applied for processing and validating the sampling and analytical laboratory data from the U.S. National Park Service (NPS) IMPROVE network. Data processing and validation for IMPROVE are the responsibility of the Data & Reporting Group within the Air Quality Research Center (AQRC) at University of California, Davis (UCD); the AQRC Data & Reporting Manager supervises the project.

This SOP covers the steps involved in receiving the sampling and analytical laboratory data, processing the flow data and conducting validation, processing the analysis data into a format suitable for further review and conducting validation, and submitting the data to Cooperative Institute for Research in the Atmosphere (CIRA)/Federal Land Manager Environmental Database (FED), the EPA's Air Quality System (AQS) database, and UCD's CSN/IMPROVE Archive (CIA) database.

This document is intended to give only the outline of how data are processed, validated, and delivered. Each of the required steps involved has a specific function and a set of procedures associated with that function. A detailed explanation of each of these steps is required. Thus, descriptions of the individual procedures are given in the Technical Instruction (TI) documents that are referenced within this SOP.

2. SUMMARY OF THE METHOD

Filter samples are collected routinely every third day throughout the year in the IMPROVE network, resulting in approximately 20,000 annual samples per module and approximately 80,000 total filters collected per year. Each site has four routine modules collecting deposit on PTFE, nylon, or quartz filters; PTFE filters are used in two of the modules, nylon and quartz filters are used in each of the other two modules. In addition, one site has a full suite of collocated modules and 13 sites have one additional collocated module.

Filter boxes are prepared by the Sample Handling Lab at the University of California, Davis (UCD) and sent to the field, where field sampling is conducted by local operators. Once the samples are received back at the UCD Sample Handling Lab after sampling, the exposed filters are sent to the laboratories at UCD, RTI International (RTI), and Desert Research Institute (DRI), along with associated operational sampling data such as sampling dates and site information.

PTFE samples are analyzed at UCD for $PM_{2.5}$ and PM_{10} gravimetric mass, elements by energy dispersive X-ray fluorescence (EDXRF), and optical absorption by Hybrid

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

Integrating Plate/Sphere (HIPS). Nylon samples are analyzed at RTI for ions by ion chromatography (IC) and quartz samples are analyzed at DRI by thermal optical analysis (TOA). Following laboratory analysis, all analytical results are assembled by UCD for processing and initial validation. Ion analysis results from RTI and carbon analysis results from DRI are received in data files, typically delivered as .csv files for ions data and .xml files from DRI, and ingested into the UCD IMPROVE database using the UCD IMPROVE Data Management website. Gravimetric mass, elemental, carbon, and optical absorption analysis results from UCD are automatically ingested.

Data processing involves calculating sample volume from field data on flow rates and sampling duration and subsequently calculating ambient concentration, uncertainty, and method detection limit (MDL) for each analyte using the laboratory result plus the sample volume. The UCD analyst will use functions in the *crocker* software package to calculate final results and post them to the UCD IMPROVE database. The analyst will also review any output messages for errors. The calculated concentrations undergo validation for technical acceptability and reasonableness based on information such as routine quality control (QC) sample results, data quality indicator calculations, performance evaluation samples, internal and external audits, statistical screening, internal consistency checks, and range checks. The analyst uses the UCD IMPROVE Data Management website along with custom software in the R language to perform validation; the primary review tools are summary data tables and comparison figures.

Once the data have been processed and validated, the analyst prepares delivery files of the validated data sets using custom tools in the *crocker* R package. The final data files are checked for correctness and then submitted to the Environmental Protection Agency's (EPA) AQS Database, the Cooperative Institute for Research in the Atmosphere (CIRA) Database (FED), and ingested into the CIA Database.

3. DEFINITIONS

- **AQRC:** Air Quality Research Center.
- **AQS:** EPA's Air Quality System database.
- **CSN and IMPROVE Archive (CIA) Database:** A database of the complete record of CSN and IMPROVE data coupled with a web-based visualization and analysis tool.
- **Chemical Speciation Network (CSN)**: EPA's $PM_{2.5}$ sampling network, with sites located principally in urban areas.
- **CIRA:** Cooperative Institute for Research in the Atmosphere.
- **crocker:** A custom software package in the R language that contains the data processing code used to produce, check, and post the final results.
- **CSV:** a comma-separated value file that is the common format for delivery files.
- **datvalIMPROVE:** A custom software package in the R language that contains the data validation code used to collect, compare, and flag the final results.
- **DRI:** Desert Research Institute.

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

- **Energy Dispersive X-Ray Fluorescence (EDXRF):** An analytical technique used to determine the concentration of elements.
- **Federal Land Manager Environmental Database (FED):** a database of environmental data managed by Cooperative Institute for Research in the Atmosphere (CIRA)
- **Hybrid Integrating Plate/Sphere (HIPS):** An analytical technique for optical absorption.
- **Ion Chromatography (IC):** An analytical technique used to determine the concentration of ions.
- **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.
- **Method Detection Limit (MDL):** A lower limit of detection specific to method of analysis and reported parameter.
- **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.
- **RTI:** Research Triangle Institute, International.
- **SQL:** database management system used by AQRC.
- **Thermal Optical Analysis (TOA):** An analytical technique used to determine the concentration of carbon. Also referred to as TOR (Thermal Optical Reflectance) and TOT (Thermal Optical Transmittance).
- **UCD:** University of CA—Davis.
- **Extensible Markup Language (XML):** a markup language defining a set of rules for encoding documents in a particular format; used for IMPROVE carbon files.

4. HEALTH AND SAFETY WARNINGS

Not applicable.

5. CAUTIONS

Not applicable.

6. INTERFERENCES

Not applicable.

7. PERSONNEL QUALIFICATIONS

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

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

7.1 Data & Reporting Group Manager

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

7.2 Lead Quality Assurance Officer

The lead quality assurance officer:

- devises techniques that improve the efficiency, traceability, and accuracy of the data management;
- develops validation criteria, automated and manual checks, and visualization tools for assessing data quality and consistency;
- reviews method detection limit (MDL) and uncertainty;
- 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.

7.3 Quality Assurance Officer

The quality assurance officer:

- 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;
- 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;
- communicates with laboratories regarding analytical issues and/or reanalysis requests;
- 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.

8. EQUIPMENT AND SUPPLIES

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.

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

- Flow rates a raw flow readings are either acquired from sampler flashcards and stored in the *sampler.FlowSourceData* table (for V2 controllers) or uploaded daily by the controller and stored in the sampler.FlowSourceDataV2 table (for V4 controllers). In addition, handwritten log sheets that contain flow readings and other sampling information recorded by the operator are stored in the *filter.Filters* and *filter.SampleCartridges* tables.
- Average flow rates (24-hour average) are calculated using a SQL procedure called *sampler.spFilterAverageFlowRates* 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 are acquired in the UCD Sample Handling Laboratory and stored in the *grav.SampleAnalysis* table.
- Carbon analysis results are acquired from files generated by Desert Research Institute (DRI; Reno, NV) TOA Laboratory and are stored in the *dricarbon.MassLoadings*, *dricarbon.CarbonLaser*, and *dricarbon.SampleAnalysis* tables.
- Ions analysis results are acquired from files generated by RTI International (Research Triangle Park, NC) IC Laboratory and are stored in the *ions.MassLoadings and ions.SampleAnalysis* tables.
- Elements analysis results are acquired from the UCD XRF Laboratory through a custom ingestion process and are stored in two tables in the database: *XRF.SampleAnalysis* and *XRF.DeviceCounts*. These are the main tables with mass loading results, reported as raw areal densities from the XRF instruments (ug/cm2). The *DeviceCounts* table contains the XRF results for each element. The *SampleAnalysis* table contains information about the filter analyzed, the instrument used for analysis, and the date and time of analysis.
- Optical absorption analysis results are acquired from the UCD Hybrid Integrating Plate/Sphere (HIPS) Laboratory through a custom ingestion process and are stored in the *hips.Results* and *hips.SampleAnalysis* tables.

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 9.3.3).

IMPROVE Management Website (https://improve.aqrc.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 pertinent to data processing and validation include:

- The Filter Section (https://improve.aqrc.ucdavis.edu/Filters) consists of web pages for searching for specific filters, reviewing operational and analytical data associated with a filter, or applying flags and comments.

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

- The Samplers Section (https://improve.aqrc.ucdavis.edu/Samplers) provides details of all IMPROVE samplers, both active and inactive sites, with options to edit information as well as options to add new samplers.
- The XRF Section (https://improve.aqrc.ucdavis.edu/Xrf/Home) is an interface for processing XRF elemental mass loadings, managing processed sets, and applying flags.
- The Analysis Data Section (https://improve.aqrc.ucdavis.edu/AnalysisData/Home) consists of web pages for importing and viewing carbon and ions data viewing mass and optical absorption data, and reviewing information on analysis pathways. Under this home page are the following subsections:
- The Operations Section (https://improve.aqrc.ucdavis.edu/Operations/Home) is a live display of the sampler status for the sites equipped with the V4 controllers. This section also consists of web pages for scheduling boxes and reviewing box details.
- The Reports Section (https://improve.aqrc.ucdavis.edu/Home/Reports) has links for IMPROVE status pages (https://shiny.aqrc.ucdavis.edu/ImproveStatus/) and IMPROVE data exploration pages (https://shiny.aqrc.ucdavis.edu/ImproveData/).

Flow Graphs (https://shiny.aqrc.ucdavis.edu/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 (https://shiny.aqrc.ucdavis.edu/ImproveData/): This web application provides interactive visualizations of processed concentrations, uncertainties, and MDLs, plus custom tools for validation as described in Section 9.3.

9. PROCEDURAL STEPS

UCD IMPROVE data processing and validation occurs in several steps, outlined below. The specifics of each step are detailed in the noted Technical Information documents.

1. Data ingest (IMPROVE TI 351A): Gravimetric mass, EDXRF, and HIPS analysis results are transferred into the UCD database through an automated service. IC analysis results files from RTI and TOA analysis results files from DRI are received via email, and results are ingested to the UCD CSN database.
2. Data processing (IMPROVE TI 351B): Operational information from field sampling and laboratory analysis results are combined to calculate concentrations, uncertainties, and method detection limits.
3. Validation (IMPROVE TI 351C): Data and metadata are reviewed through a variety of visualizations and summary data tables. Several statistical and visual checks are applied and examined. Reanalyses are requested as needed. Data are flagged with informational and/or flags (status) as appropriate.
4. Data delivery (IMPROVE TI 351D): Data are formatted into two formats for delivery to the CIRA/FED database and another format for delivery to the AQS and CIA databases.

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

5. Flow validation (IMPROVE TI 351E): Flow data from the network are reviewed and validated using various tools including a flow plotter website.
6. Data Preparation and Reporting (IMPROVE TI 351F): Box information is created or modified in the UCD database. New sampling site metadata is also added to the UCD database and AQS database. Quarterly site sampling information is compiled into a report and delivered to IMPROVE related personnel.

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 Data Disaster Recovery Plan

The scope of data 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.

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

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.

DocuSign Envelope ID: D2A7C4FB-A926-4DD1-ACCD-8F4006E289EF

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 SOP.

12. 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.; Lioy, P.J.; Mueller, P.K. (1995). The measurement process: Precision, accuracy, and validity. In Air Sampling Instruments for Evaluation of Atmospheric Contaminants, 8th; American Conference of Governmental Industrial Hygienists: Cincinnati, OH, 187-194.