# Data Flow, Application of Data Quality Flags, and Data Validation Processes for CCAQS

## 1.0 Introduction

The success of an air quality study is dependent on the credibility of the data collected and the controlled processes used to establish data quality. If controlled processes are not in place, the accuracy of the data may be questioned. The data researchers need to be assured of the integrity of the processes performed during the creation of the final data set for a study. For this purpose, the data flow and validation processes for the Central California Air Quality Studies (CCAQS) have been developed. This document describes these processes, descriptions of the data flow, and validation processes used to identify, track and maintain the quality of CCAQS data. It covers data flow as data are collected, processed, stored and reviewed in the CCAQS Database System. By clearly defining these processes, the credibility of the numeric data are supported at every step.

The basic framework for all CCAQS data processing utilizes four-levels of data quality assurance (QA) designation: Level 0; and Levels 1, 2, and 3 as defined and applied by Mueller (1980, 1983), Watson (1993, 1989, 1993, 1995), Tombach (1987, 1996), Korc (1996) and Chow (1998). Associated with these data quality levels is a comprehensive set of data quality flags. These flags provide data value specific quality indicators. This information is invaluable to data analysts and air quality modelers. The "flagging" of data provide a source of "documentation" supporting the process of establishing data validity and maintaining the credibility of the data for air quality modeling and other analytical purposes.

This document describes how the four QA levels 0-3 are utilized and outlines the QA data flow and validation processes. It also serves as a reference document for data providers during their review of data that they have collected and analyzed. This document also describes what processing occurs after the data are received and entered into the database system, thereby serving as a reference for data analysts and modelers who use the data.

## 2.0 Shared Responsibilities

In the data flow process, data providers (contractors) and the data manager are the primary people responsible for maintaining the integrity of the entire process and validity of the data. The various individuals that audit analytical instruments and related processes also perform an integral role. The air quality modelers, data analysts and researchers provide invaluable input for establishing levels of data quality assurance. This is particularly important after the data has been collected and is in the database.

### 2.1 Data Collection and Initial Processing

The documentation to support the credibility of data collection and initial data quality assurance are the responsibility of the data provider. This is true up to the point where the data are loaded into the CCAQS Database System. This includes the process of data collection, application of calibration factors, initial QA, data analysis, data "flagging", rollups (averaging) and reporting. A combination of data record notes, data quality flags and process documentation are all part of this first phase of processing. During the data collection phase, one role of the data provider is to assist in maintaining process credibility and validity of data by contributing to the efforts of the CCAQS Data Manager in the following three ways:

1. Providing a scientific basis that the information being accessed is ***valid***.

   ***"Validity"*** is supported by documentation that provides:

   - Proof that all applicable standard scientific procedures were adhered to.

   - Precise descriptions of all collected and processed numeric data. (This is *metadata* and is defined as data that describes other data. Among many other data elements, metadata includes: collection method; instrument type; instrument accuracy; instrument precision; data format; unit conventions; variable naming conventions; QA/QC flags.

   - Justify technically all calculations and processes including, parameter interpolations and quality assurance criteria. In addition, technical justification should be provided for all scientific conclusions based upon new data processing routines. For example, this could include special data processing and analysis procedures built into the data management system. This enables inter-comparisons of new state-of-science monitoring technologies with existing technologies.

   - Reference external information upon which calculations, processes and conclusions are based.

2. Allowing all procedures performed in creating a data set at Level 1A to be ***traceable***.

   ***"Traceability"*** is a documented history of all processes performed on each raw data set transmitted to the database. Traceability is assured by maintaining tabulated, chronological listings, which summarize each step that is performed along with the method by which it is performed. It indicates how this occurred (e.g., program name, etc.) along with the verification and quality assurance procedures implemented and the corresponding results.

**3.** Ensuring that results produced during a study, including the final data set, are ***reproducible***.

   ***"Reproducibility"*** allows the duplication of results from any data validity level. Reproducibility requires traceability, since all processing steps performed in producing specific results must be duplicated. Reproducibility requires that all data

management tools used be stored together with a chronological set of data validation records for all data sets (e.g., source code for processing programs must be stored and available if needed).

It is of primary importance that data providers and the data manager make the necessary efforts to ensure that all aspects of the data collection, handling, and analysis and evaluation are well documented. This is essential with respect to considerations of data validity, traceability, and reproducibility. Documentation accompanying data are a requisite for providing a data *history*, which gives value to the data. To accomplish this requires that good reporting procedures be maintained and implemented at each step of data handling and processing.

## 2.2 Data Management and Quality Assurance

Following the initial data collection and review by the data providers, the CCAQS data files are transmitted via ftp to an ftp server where they can be loaded into the CCAQS Database System. The "loading" process utilizes an automated file screening application, which checks for correct data file formatting per the Central California Air Quality Studies (CCAQS DTFD). It also checks to ensure that only the correct data elements from the CCAQS Reference Tables have been included. The screening process includes air quality and meteorological parameter range-checks and generates a short report based upon the screening outcome. The results are reported as descriptive error messages sent as emails to the data provider as well as the CCAQS data manager. After the screening and data loading the data are accepted and flagged internally as Level 1A data. Once all issues are resolved regarding a data file submittal and file screening, the subsequent data QA processing and associated documentation becomes the responsibility of the CCAQS Data Manager. In the data flow process, implementing the next level of data quality beyond Level 1A, is the responsibility of the data manager. This will require input from data analyst and modelers. With the assistance of automated QA software applications and input from various data researchers, the data can attain Level 1B, Level 2 or Level 3 status within the database system.

The data manager will use a relational database management system (RDBMS) to provide the data management and implementation of the subsequent levels of quality assurance and data validation. Utilizing a data management system will help maintain the quality of the data and the credibility of the studies. The system will facilitate linking critical information such as raw data, processed data, processing programs, and quality assurance programs.

# 3.0 Maintaining Instrument and System Performance

## 3.1 Calibration Values and Uncertainty Estimates

Data contractors and laboratory personnel periodically calibrate each instrument involved in the data collection/analysis processes. Routine instrument calibration provides the calibration factors that are applied to data values. The application of calibration factors to collected air quality data is an integral part in determining the actual concentration of the parameters being monitored. During this process, data flagging may be applied to indicate that data were associated with a failed calibration sample, etc. Calibration values for all measurements are reported as a separate (ancillary) file using the same

reporting conventions and formats as the actual measurement. Access to calibration values is crucial for quality-assurance, and to some extent for data analysis and modeling.

Uncertainty estimates should be reported for each parameter. This is very important particularly in cases where air-monitoring equipment is being evaluated. Uncertainty estimates should always provided along with the observation data values. There are many factors that can affect uncertainty, which are usually found in the SOP (Standard Operating Procedures) developed for an air quality study. The SOP is where the method used by the investigator to calculate uncertainty for each parameter should be defined.

## 3.2 Audit Program

The principle objective of the CCAQS Audit Program is to ensure that the established data quality objectives (DQOs) are achieved. Usually, these DQOs are reviewed first in the applicable Quality Integration Work Plans (QIWP) and standard operating procedures (SOPs) for the measurement systems. This ensures that the DQOs are adequate and realistic for the intended purpose and that the proposed procedures are appropriate to achieve those goals. Testing the procedures documented in the SOPs. This involves conducting system and performance audits. A “systems” audit is a qualitative evaluation of the overall operations of an analytical laboratory. Emphasis is given to confirming that all project quality control (QC) procedures are performed and that acceptance criteria are met. Preventive maintenance is evaluated as well as corrective action procedures. Randomly selecting a reported value and working back through intermediate calculations and supporting paper work to the raw values can be done to verify data traceability. This type of checking confirms error free calculation and ensures that supporting documentation is present.

For the CCAQS, the general approach for auditing is to assure sure that the following items are addressed:

- Air quality and surface meteorological measurements receive system and performance audits at the beginning of the annual study, as well as a second performance audit immediately prior to the winter intensive study (approximately one year later). The second set of performance audits provides sufficient opportunity to identify any significant system problems.

- Any air quality or surface meteorological measurements specific to the intensive studies will receive system and performance audits immediately prior to an intensive operating period (IOP).

- The upper air meteorological measurement will receive systems performance and system audits at the beginning of the Study. Any upper air meteorological measurement systems specific to the intensive studies receive systems and performance audits just prior to an IOP.

- Laboratories receive system and performance audits within the first three months of the start of the study.

- Data processing efforts receive system audits within the first six months of the start of the annual study.

The auditing procedures described above provide an independent assessment of the performance for measuring systems. Auditing of measurement equipment ensures that the instrument providing data values can be relied upon to provide valid data.

## 4.0 Quality Assurance Applications

The CCAQS Database System will include a table to maintain the internal database QA routines. This table will include the following:

1. Version_ID
2. Code_Name
3. Code (memo field containing the code)
4. Date_Coded
5. Purpose
6. Programmer
7. QA_Date
8. QA_Person

This will allow all Level 2A data QA codes to be 'extracted' from the database, per a query including the latest version identification (Version_ID). It will allow easy rollback and traceability to earlier QA code versions. Similarly, there will be a database identifying which versions of QA codes have been applied to which data files for traceability purposes.

The QA applications are being developed using Microsoft Visual Basic code. Within the QA code files there will be code description and identification and version maintenance. This will be included in header file for each module. This will include the Items 1-8 above along with what is normally found in a header file such as defined Variables, Variable_Descriptions, and variable naming conventions.

The data manager is to assure that only "valid" data are accessible to data users. This is an integral part of the internal controls that incorporated during the development of the database. Suspect and Invalid data are not available for download but can be reviewed. This ensures that different modelers and data analysts do not misinterpret or inadvertently misuse data "known" to be bad.

## 5.0 Components of the Data Flow Process

### 5.1 Data Quality Flags

There are a variety of situations where data flags can be applied. The actual number of flags available is dependent on the extent of specific information that needs to be communicated by those submitting the data. Details of the flags reflect the desire of those that will be using the data. Data flag codes should

differentiate between valid values, invalid values, estimated values, interpolated values, minimum detectable limits (MDL) values, and missing values, etc.

Another important consideration is the averaging times of raw data into quality assured data. This is the data averaging that occurs as part of obtaining the final value used by air quality modelers and data analysts. For example, there are 5-minute data values that are averaged to obtain 1-hour averages. There may be one or multiple steps involved in averaging study data values. For instance, data collected every minute (60 data points total) may be averaged to obtain a representative hourly average value. Alternatively, observations collected every minute (5 data points total) may be averaged to representative five-minute averages, which are then averaged into one hour averaged values. This results in data that can be categorized as either post- or pre-averaged data. At each step prior to roll-up, data provider can apply their own data flags. A worst-case example might be where data is rolled-up multiple times from one-minute raw data to five-minute averaged value and then again to an hourly average. The one-minute raw data values are first "flagged", and then rolled-up to a five minute averaged value, which are then flagged again. Averaging occurs one more time to obtain hourly averages and again new flags are assigned to these data values. Often the flags associated with the raw 1-minute data and the 5-minute averaged data are discarded entirely and lost in the data flow process. This can be unfortunate later when modeler or data analysts need traceability back to the raw data. Ideally, the data flow process should maintain a history of all of the data flagging that occurs. Currently, there is no mechanisms set up to do this.

For supplemental data (data sets not evaluated by the processing established within the study) the quality assurance flags will be obtained from the system supplying these data. The use of flags is often limited to a valid or invalid flag only. Users of the data should pay close attention to the data validation procedures and designations used by those networks (AIRS, CIMIS, RAWS) before using the data for a particular research application.

### 5.1.1 Flag Class Definitions

Data flags can be applied during the sampling, measurement, analysis and validation stages of the CCAQS data flow process. The flags are organized into four general classes within CCAQS. These are the Activity Flags, Primary Flags, Secondary Flags, and Data Validation flags. They are applied during "process points" through the data flow. As data progresses through the sequence of processing steps, the pertinent flag(s) are associated with individual data points. Which flags are applied is dependent on the specific circumstances surrounding each data point. Flags remain associated with the data points and as a result carry additional data related to information defined for each flag. The four CCAQS classes of flags are defined as follows:

***Activity Flags –*** identify an "exceptional event" that could have affected the data value (e.g., forest fire).

***Primary Flag –*** these provide a high level assessment of the validity/quality of a data value. Each Primary Flag is associated with a specific set of Secondary Flags.

***Secondary Flags –*** these provide more detail and are used to identify a condition affecting data collected as either a direct/in situ/continuous field measurement or laboratory measurement on a

sample and related to the defined procedures for the quantitative analysis. They can identify a characteristic affecting a sample or any sample derived data values as well.

***Data Validation Flags –*** used as post data collection (post sampling and measurement) flags. These flags are assigned to measured values based on statistical factors and observed anomalies inherent in the data itself that were identified as a result of an analytical data validation process.

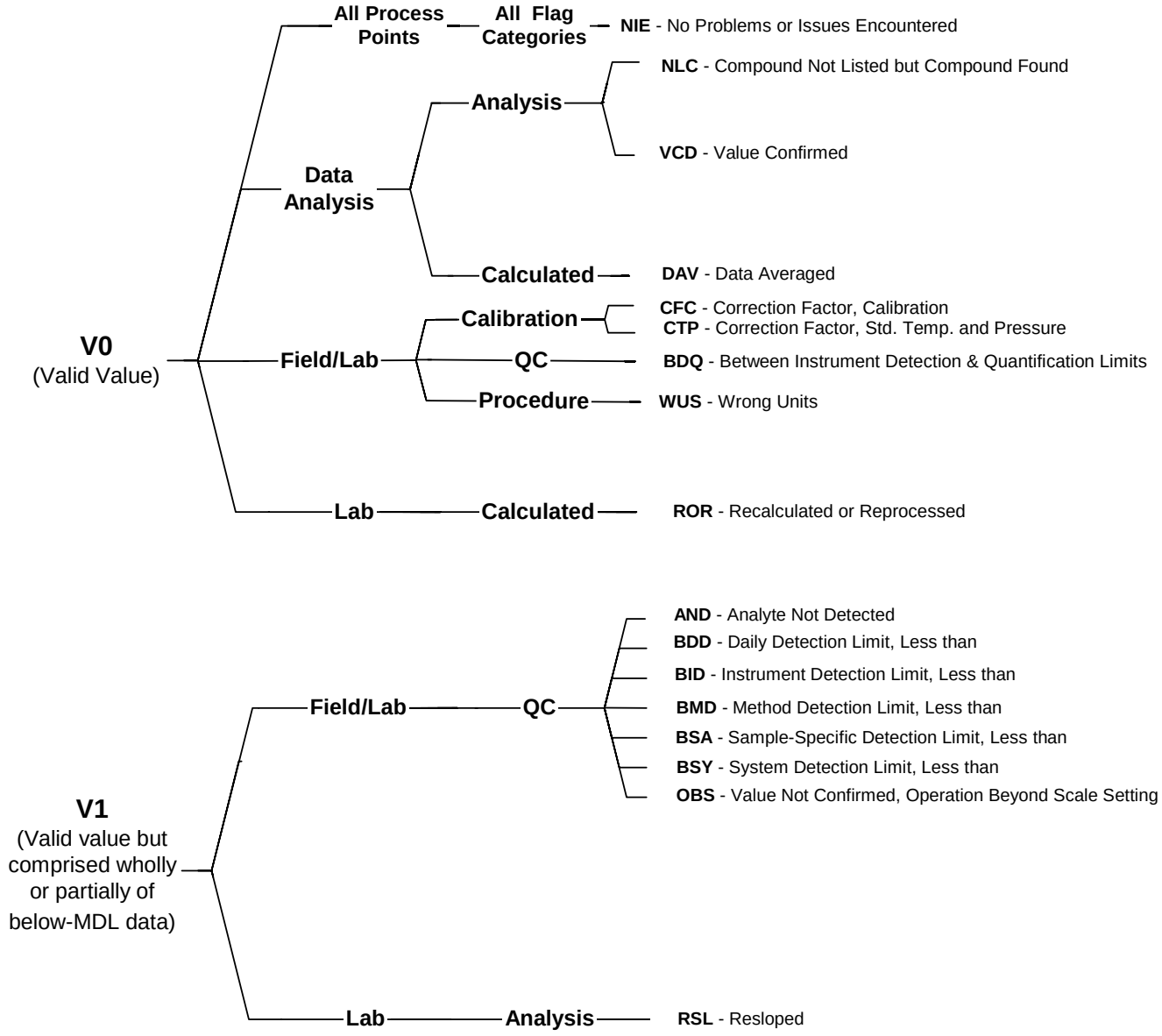
## 5.2 Flag Classifications

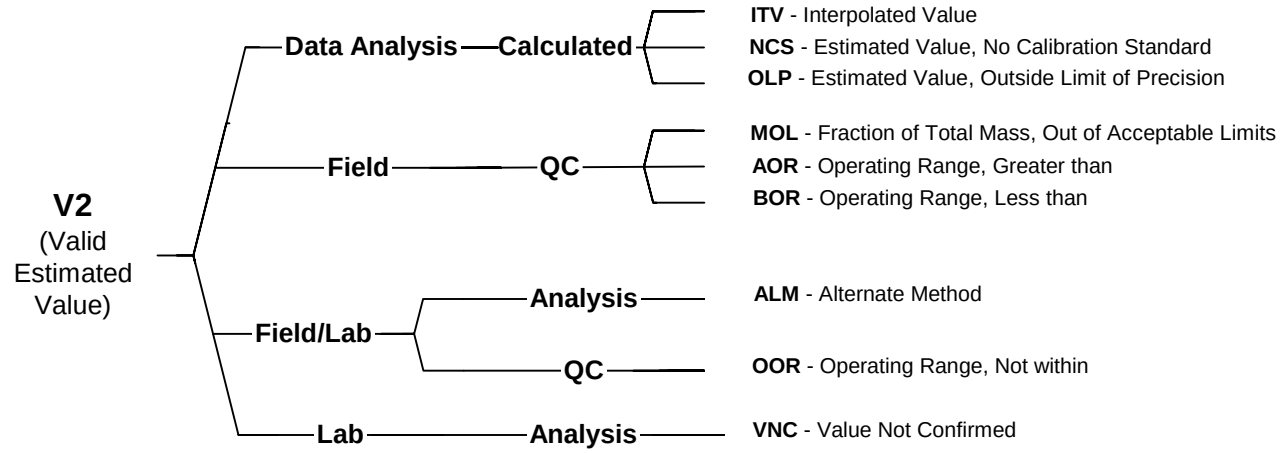
The CCAQS Data Transmittal Format Document (CCAQS DTFD) was developed to accommodate Primary, Secondary and Activity Flags. Referring to Table 1, Study Flags and Processes, the *Primary Flag* indicates the "status" of the data. There are three variations of primary valid flags: V0, V1, V2. The other three are S (suspect), M (missing), or I (invalid). The use of a primary flag can result from conditions encountered during field sampling, field or lab measurement or the initial data analysis phase. They carry the highest-level of quality designation in the hierarchy of quality assurance. The *Secondary Flag* is used to reflect the most important sampling, measurement or analysis consideration affecting the data. These can be seen in found in Diagrams 1 below. They provide more detailed information pertinent to the data. The data manager can add other flags to these lists as needed. The *Activity Flag* is used to identify exceptional events or certain environmental conditions that could have influenced the data monitoring in the field. These flags are listed in Diagram 2 below.

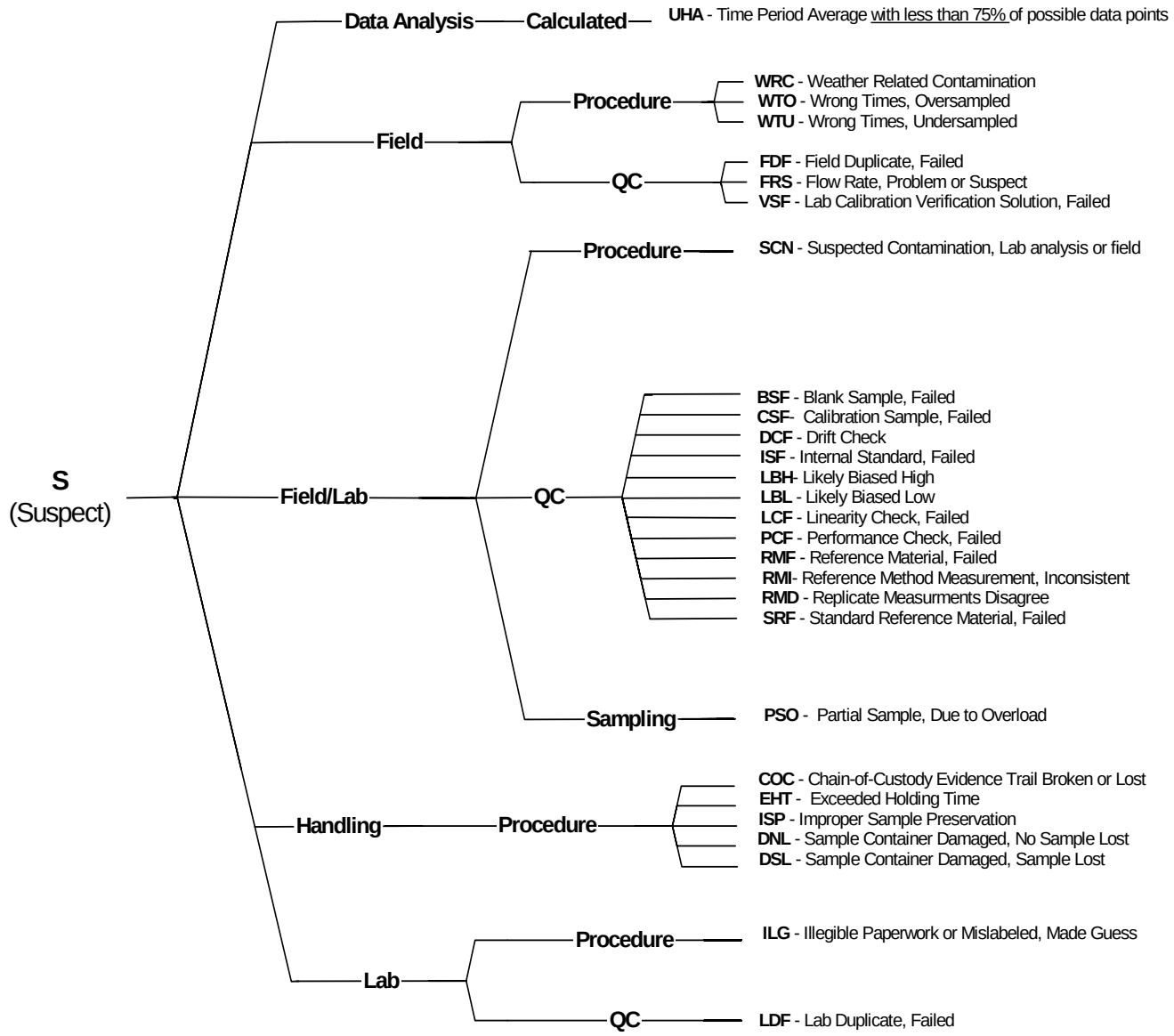
### TABLE 1. Study Flags and Processes

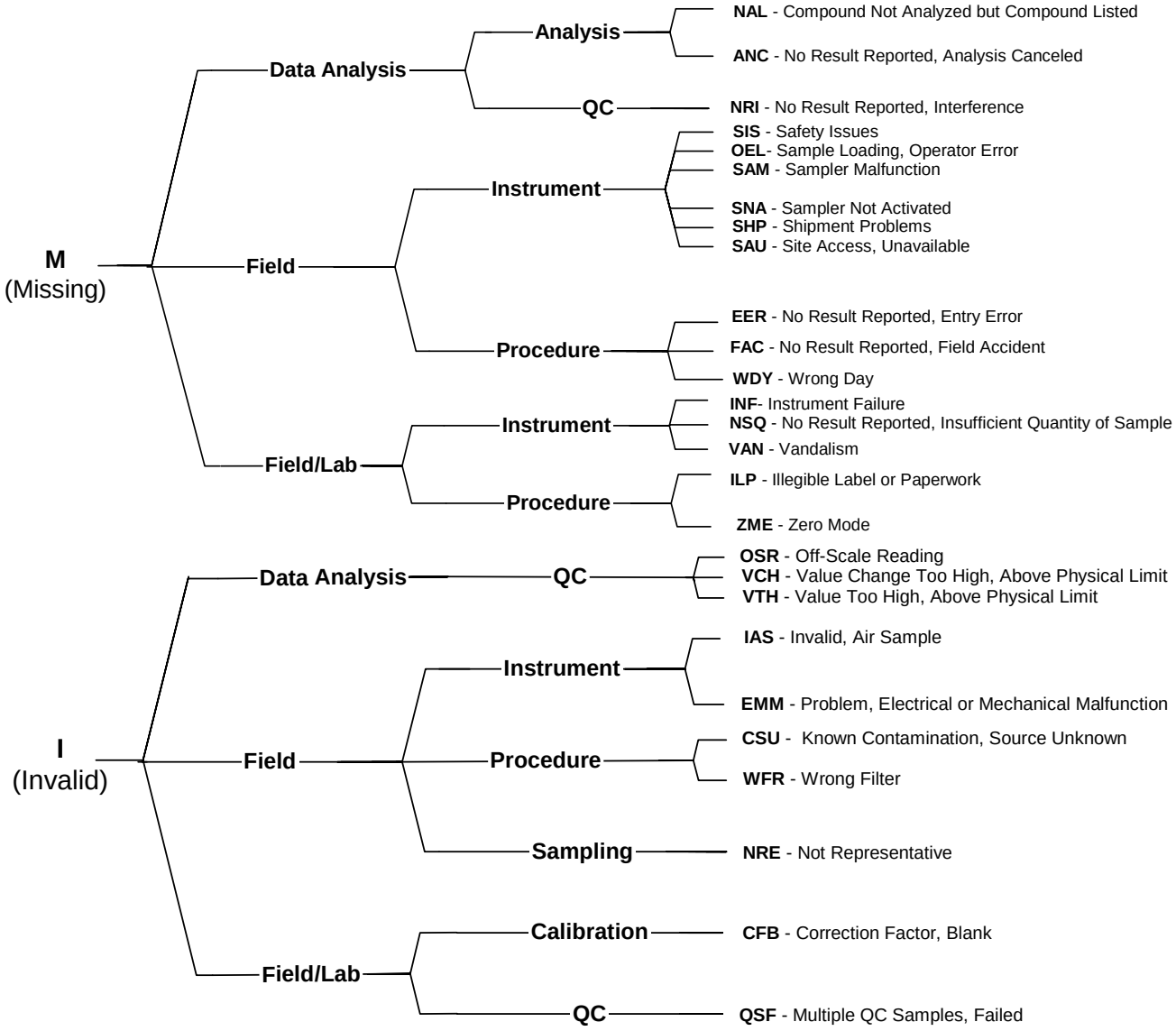
| Process Point | Primary Flag | Secondary Flag | Activity Flag | QA Level | Internal CCAQS DB Flags |
|---|---|---|---|---|---|
| **Field/Lab/Data Analysis** | V0, V1, V2, S, M or I. (See Diagram 2 below) | Detailed flag. Dependent on the Selection of the Primary Flag. (See Diagram 2 below) | Exceptional Event or Environmental Condition Flag (See Diagram 1 below) | 0 - 1A | NA |
| **Data Validation** (post-sampling & measurement analysis) | NA | NA | NA | 1A, 1B, 2, 3 | TBD |

# DIAGRAM 1. Primary & Secondary Flags

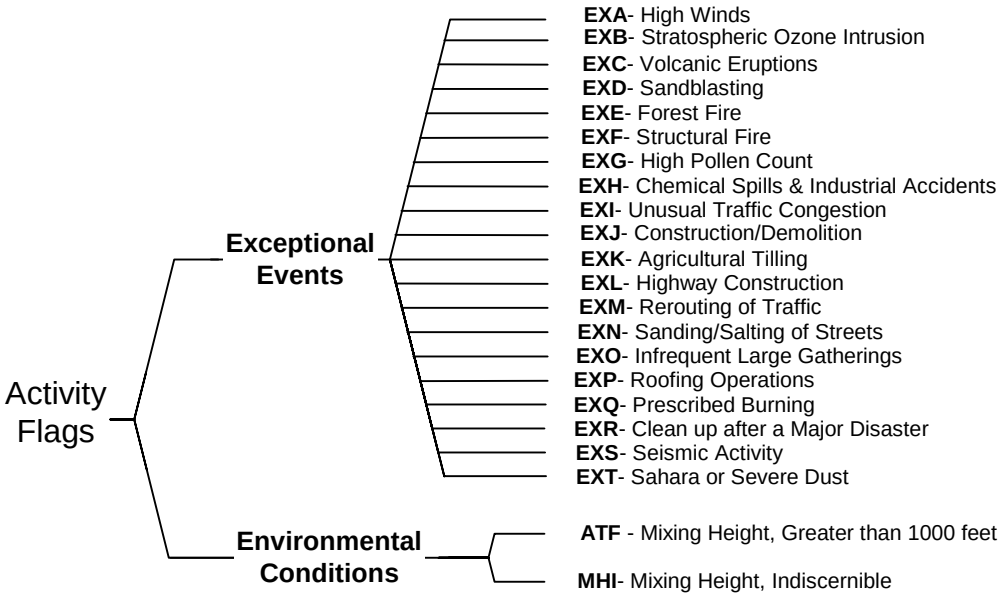



V2
(Valid Estimated Value)
Data Analysis
Calculated
ITV - Interpolated Value
NCS - Estimated Value, No Calibration Standard
OLP - Estimated Value, Outside Limit of Precision
Field
QC
MOL - Fraction of Total Mass, Out of Acceptable Limits
AOR - Operating Range, Greater than
BOR - Operating Range, Less than
Field/Lab
Analysis
ALM - Alternate Method
QC
OOR - Operating Range, Not within
Lab
Analysis
VNC - Value Not Confirmed


S
(Suspect)
Data Analysis
Calculated
UHA - Time Period Average with less than 75% of possible data points
Field
Procedure
WRC - Weather Related Contamination
WTO - Wrong Times, Oversampled
WTU - Wrong Times, Undersampled
QC
FDF - Field Duplicate, Failed
FRS - Flow Rate, Problem or Suspect
VSF - Lab Calibration Verification Solution, Failed
Field/Lab
Procedure
SCN - Suspected Contamination, Lab analysis or field
QC
BSF - Blank Sample, Failed
CSF- Calibration Sample, Failed
DCF - Drift Check
ISF - Internal Standard, Failed
LBH- Likely Biased High
LBL - Likely Biased Low
LCF - Linearity Check, Failed
PCF - Performance Check, Failed
RMF - Reference Material, Failed
RMI- Reference Method Measurement, Inconsistent
RMD - Replicate Measurments Disagree
SRF - Standard Reference Material, Failed
Sampling
PSO - Partial Sample, Due to Overload
Handling
Procedure
COC - Chain-of-Custody Evidence Trail Broken or Lost
EHT - Exceeded Holding Time
ISP - Improper Sample Preservation
DNL - Sample Container Damaged, No Sample Lost
DSL - Sample Container Damaged, Sample Lost
Lab
Procedure
ILG - Illegible Paperwork or Mislabeled, Made Guess
QC
LDF - Lab Duplicate, Failed


M
(Missing)
Data Analysis
Analysis
NAL - Compound Not Analyzed but Compound Listed
ANC - No Result Reported, Analysis Canceled
QC
NRI - No Result Reported, Interference
Field
Instrument
SIS - Safety Issues
OEL- Sample Loading, Operator Error
SAM - Sampler Malfunction
SNA - Sampler Not Activated
SHP - Shipment Problems
SAU - Site Access, Unavailable
Procedure
EER - No Result Reported, Entry Error
FAC - No Result Reported, Field Accident
WDY - Wrong Day
Field/Lab
Instrument
INF- Instrument Failure
NSQ - No Result Reported, Insufficient Quantity of Sample
VAN - Vandalism
Procedure
ILP - Illegible Label or Paperwork
ZME - Zero Mode
I
(Invalid)
Data Analysis
QC
OSR - Off-Scale Reading
VCH - Value Change Too High, Above Physical Limit
VTH - Value Too High, Above Physical Limit
Field
Instrument
IAS - Invalid, Air Sample
EMM - Problem, Electrical or Mechanical Malfunction
Procedure
CSU - Known Contamination, Source Unknown
WFR - Wrong Filter
Sampling
NRE - Not Representative
Field/Lab
Calibration
CFB - Correction Factor, Blank
QC
QSF - Multiple QC Samples, Failed

## DIAGRAM 2. Activity Flags



# 6.0 Data Flow Overview

[**Insert** – Data Flow Diagram with "steps" used below to reference. Add additional discussion as needed.]

# 7.0 Data Flow: A Step-By-Step Example

## 7.1 Processing Level 0 Data

Step 0.1 Obtain either raw data observational values directly from instrument readouts or from the analysis of air quality samples from air quality sampling equipment. The "Level 0" data has been defined as below. Additional considerations for CCAQS are also indicated below:

> **Level 0 (0):** These data are obtained directly from the data loggers that acquire data in the field. Averaging times represent the minimum intervals recorded by the data logger, which do not necessarily correspond to the averaging periods specified for the database files. Level 0 data have not been edited for instrument downtime, nor have procedural adjustments for baseline and span changes been applied. Level 0 data are not contained in the database, although they are consulted [evaluated] on a regular basis to ascertain instrument functionality and to identify potential episodes prior to receipt of Level 1A data.
>
> [**Level 0 for CCAQS**: In addition to data loggers, CCAQS data are also derived from the collection of air samples that are sent to the laboratory for analyses. Level 0 data are not contained in the CCAQS database. At a later time the resultant "raw" Level 0 data values may be averaged (e.g., 1 minute data averaged to 5 minute data values, 5 minute data averaged to 1 hour data values). Data reported in the CCAQS database will often be an averaged value and therefore reported at a different sampling frequency from the raw data. This is also true for "supplemental" data like that from AIRS, RAWS, etc. ]

Step 0.2 Provide log of field events and pertinent factors and other circumstances that impact the data observation value(s) or air samples . These factors can be independent of, or prior to, during or after obtaining an actual measurement. Assess surrounding activities that may have impacted monitoring and the resultant data values.

Step 0.3 Translate your information and flagging scheme from the data sample collection and measurement Flag raw data accordingly, using a Primary Flag (Refer to Diagram 1). Also, apply an Activity Flag if an exceptional event was encountered (Refer to Diagram 2. Activity Flags).

Step 0.4 "Validate" measured data values and air samples based on the circumstances discovered surrounding the data. Associate the data or samples with one of the Primary Flags: V0, V1, V2, S, M or I. If data are to be derived from air samples continue to Step 0.5, otherwise go to Step 0.6.

Step 0.5 Analyze samples in the laboratory and obtain data observation values. Log lab related events, such as sample handling and other pertinent factors and circumstances that impact the data observation value(s) in the laboratory. Apply the appropriate Primary and Secondary Flag (Refer to Diagram 1). Continue with Step 7.

Step 0.6 Validate data observation values derived from the field measurement or laboratory analysis based on the severity of the flag(s). Flag data as "V0, V1, V2, S, M or I".

Step 0.7 Determine the instrument calibration factors and apply calibration factors to the data observation values. Do not replace invalidated data with –99 or anything similar, but instead maintain the invalidated data values within the data file.

Step 0.8 Average the pre-rolled up raw data and apply the appropriate flags to the "derived" data. When assigning new data flags to the data obtained by averaging or "rolling-up" data, maintain the flags and data values for the data used in the roll up within your database or raw data file. Accompany the averaged data with notes indicating how new flags were assigned based on the flags for the pre-rolled up raw data. This will provide the traceability in the CCAQS Database System.

## 8.0 Establishing Data QA Level 1A & 1B Data

Step 1.0 Closer data analysis and review begins as the data is moved from Level 0 to Level 1A. It is important that data providers identify the specific steps in the data processing. The steps specific to a data set should be documented electronically, along with the data such that the result(s) from each step can be readily referenced. This will help minimize problems resulting from data related questions that arise months or years later. This will also help avoid the situation where the responsible individual leaves for other employment and cannot be contacted or does not remember the specific circumstances surrounding a data set. At the very beginning this information should be available electronically to researchers that may request it.

Apply the Level 1A criteria defined as follows:

***Level 1A (1A):*** These data have passed several validation tests applied by the measurement investigator prior to data submission. The general features of

1. ~~Removal of data values and replacement with -99 when monitoring instruments did not function within procedural tolerances~~; Retain all data values and use the appropriate flags. Missing values should be 'null' and flagged as missing.

2. Flagging measurements when significant deviations from measurement assumptions have occurred;

3. Verifying computer file entries against data sheets;

4. Replacement of data from a backup data acquisition system in the event of failure of the primary system;

5. Adjustment of measurement values for quantifiable baseline and span or interference biases; and

6. Identification, investigation, and flagging of data that are beyond reasonable bounds or that are unrepresentative of the variable being measured (e.g. high light scattering associated with adverse weather).

**Level 1A for CCAQS**: In the CCAQS Database System no removal of invalidated or missing data occurs. Instead the use of data flagging is specified when data are missing or invalid, i.e., there is no replacement of missing values or invalid values with –99.

After data analysis and QA processes are completed, data are identified as Level 1A. The important consideration for data validation, during the process of elevating data from Level 0 to Level 1A, is that there be traceability throughout the process. Also, an explanation of the data processing and quality assurance (QA) that were utilized from Level 0 (raw data) to Level 1A will be needed. This includes a detailed description of the data review/analysis flow used by the data provider to establish Level 1A.

Step 1.1 Data files are transmitted to the CCAQS Data Manager using the format described in the CCAQS Data Format Transmittal Document. Files are screened and a report is generated that is emailed to the data provider and the study data manager. Data are considered at Level 1A when it passes the screening process and is loaded into the database.

[**Insert** - Discussion of database facilities to support processing at Level 1A. Add additional steps as appropriate. Include discussion of "Supplemental" data processing and differences in flagging. Define flagging post- and pre-rolled up data and describe application of flags in this situation.]

Step 1.2 Apply Level 1B criteria as defined below:

***Level 1B (1B):*** Pre-programmed consistency and reasonability tests are applied by the data manager prior to integration into the CCAQS Database System. Consistency tests verify that file naming conventions, data formats, site codes, variable names, reporting units, validation flags, and missing value codes are consistent with project conventions. Discrepancies are reported to the measurement investigator for remediation. When the received files are consistent, reasonability tests are applied that include:

1. Identification of data values outside of a specified minimum or maximum value;

2. Values that change by more than a specified amount from one sample to the next; and

3. Values that do not change over a specified period.

Data identified by these filters are individually examined and verified with the data supplier. Obvious outliers (e.g., high solar radiation at midnight, 300 °C temperature) are invalidated and a Primary Flag or

"I" is submitted as part of the data record. The bounds used in these tests will be determined in cooperation with measurement investigators and network operators.

=> [**Insert** - Discussion of database facilities to support processing at Level 1B. Add additional steps as appropriate and application of DV flags.]

## 9.0 Establishing Data QA Level 2

Step 2.0 The CCAQS Database System will be used to apply Level 2 criteria as defined below by assisting in the review and applying the appropriate data validation flags:

***Level 2 (2):*** Level 2 data validation takes place after data from various measurement methods have been assembled in the master database. Level 2 validation is the first step in data analysis. Level 2 tests involve the testing of measurement assumptions (e.g. internal nephelometer temperatures do not significantly exceed ambient temperatures), comparisons of collocated measurements (e.g. filter and continuous sulfate and absorption), and internal consistency tests (e.g. the sum of measured aerosol species does not exceed measured mass concentrations). Level 2 tests also involve the testing of measurement assumptions, comparisons of collocated measurements, and internal consistency tests.

⇨ [**Insert** - Discussion of database facilities to support processing at Level 2. Add additional steps as appropriate and application of DV flags.]

Data validation is the process of determining and denoting the quality of a data set(s). These data can have either a common method of collection or data collected by various methods in one location. The validation process consists of evaluating the internal, spatial, temporal and physical consistency of each data set for invalid data and for outliers (data that are physically, spatially, or temporally inconsistent). During validation, physically unrealistic data are invalidated and reported using the Primary Flag of "I" for invalid, biases and instrumental drift are noted, and gross errors are identified. The objective of the process is to produce an archive with values that are of known quality.

## 10.0 Establishing Data QA Level 3

Step 3.1 The CCAQS Database System will be used to apply Level 3 criteria as defined below by assisting in the review and applying the appropriate data validation flags:

***Level 3 (3):*** Level 3 is applied during the [analysis phase of the] model reconciliation process, when the results from different modeling and data analysis approaches are compared with each other and with measurements. The first assumption upon finding a measurement, which is inconsistent with physical expectations, is that the unusual value is due to a measurement error. If upon tracing the path of the measurement, nothing unusual is found, the value can be assumed to be a valid result of an environmental cause.

The Level 3 designation is applied only to those variables that have undergone this re-examination after the completion of data analysis and modeling. Level 3 validation continues for as long as the database is maintained. A higher validation level assigned to a data record indicates that those data have gone through, and passed, a greater level of scrutiny than data at a lower level. All data in the CCAQS data set will achieve Level 1B status prior to use in data analysis and modeling. The validation tests passed by Level 1B data are stringent by the standards of most air quality and meteorological networks, and few changes are made in elevating the status of a data record from Level 1B to Level 2. Since some analyses are applied to episodes rather than to all samples, some data records in a file will achieve Level 2 designation while the remaining records will remain at Level 1B. Only a few data records will be designated as Level 3 to identify that they have undergone additional investigation. Data designated as Levels 2 or 3 validations are not necessarily "better" than data designated at Level 1B. The level only signifies that they have undergone additional scrutiny as a result of the tests described above.

=> [**Insert** - Discussion of database facilities to support processing at Level 3. Add additional steps as appropriate and application of DV flags.]

## 11.0 Archiving Data to NARSTO Public Data Archive (PDA)

The final step in the data flow process is the archiving of air quality study data to the NARSTO Public Data Archive (PDA). This process is the responsibility of the CCAQS Data Manager. A translation of the data flags found in Diagram 2 above to the NARSTO flags has been made. These are also included in the diagrams. The Consensus Metadata Standard: Data Quality Flags will be used to provide the data flag information that is expected by NARSTO as part of the data archiving process.

_____

## 12.0 References

1. Chow, J.C., Zielinska, B., Watson, J., Fujita, E., Richards, W.L., Neff, W., Dietrich, D., Hering, S. *Northern Front Range Air Quality Study Final Report. Volume A: Ambient Measurements*. EPRI Report TR-113023, Colorado State University, Ft. Collins CO, http://www.nfraqs.colostate.edu, 1998, p A.7-5ff.

*2.* Korc, M., Roberts, P., Blumenthal, D., *NARSTO-Northeast Data Management Plan, Version 3.1*. EPRI Report TR-to be supplied, Electric Power Research Institute, Palo Alto CA, 1996, p2-11ff.

3. Mueller, P.K., "Comments on the Advances in the Analysis of Air Contaminants". JAPCA 30:988, 1980.

4. Mueller, P.K., Hidy, G.M., Baskett, R.L., Fung, K.K., Henry, R.C., Lavery, T.F., Nordi, N.J., Lloyd, A.C., Trasher, J.W., Warren, K.K., and Watson, J.G. *Sulfate Regional Experiment (SURE): Report of Findings*. EPRI Report EA-1901, Electric Power Research Institute, Palo Alto, CA, 1983.

*5.* Tombach, I., Allard, D. W., Drake, R. L., Lewis, R. C., *Western Regional Air Quality Studies - Visibility and Air Quality Measurements: 1981-1982*. EPRI Report EA-4903, Electric Power Research Institute, Palo Alto, CA, 1987. p 8-12ff.

*6.* Tombach, I., Seigneur, C., McDade, C., Heisler, S., *Dallas-Fort Worth Winter Haze Project, Volume 2: Measurements and Haze Climatology*. EPRI Report TR-106775-V2, Electric Power Research Institute, Palo Alto, CA, 1996. p 3-17ff.

*7.* Watson, J.G., Lioy, P.J., Mueller, P.K. *The Measurement Process: Precision, Accuracy and Validity in Air Sampling Instruments for Evaluation of Atmospheric Contaminants*. American Conference of Governmental Industrial Hygienists, Cincinnati, OH, 1983.

*8.* Watson, J.G., Lioy, P.J., Mueller, P.K. *The Measurement Process: Precision, Accuracy, and Validity*. Air Sampling Instruments for Evaluation of Atmospheric Contaminants. Cincinnati, OH, 1989.

9. Watson, J.G., Roth, P.M., Ziman, S.D., Neff, W.D., Magliano, K.L., Pederson, J.R., Solomon, P.A., and Thuiller, R.H. *Data Analysis Plan for the San Joaquin Valley Air Quality Study/Atmospheric Utilities Signatures, Predictions and Experiments (SJVAQS/AUSPEX) Program*. Report No. DRI Document 8932-001.1F1, Prepared for Pacific Gas & Electric Co., San Ramon, CA; San Joaquin Valleywide Air Pollution Study Agency, Fresno, CA; California Air Resources Board, Sacramento, CA, Desert Research Institute, Reno, NV, 1993.

10. Watson, J.G., Lioy, P.J., Mueller, P.K., "The Measurement Process: Precision, Accuracy, and Validity. In Air Sampling Instruments for Evaluation of Atmospheric Contaminants", Cohen, B., and Hering, S.V., editors. American Conference of Governmental Industrial Hygienists, Cincinnati, OH, 1995.