

APPENDIX C

DATA MANAGEMENT

C. DATA MANAGEMENT

The CCOS data base will be compiled, documented, evaluated, and distributed by the Technical Support Division of the Air Resources Board (ARB). Common data management and validation conventions need to be assembled in a consistent and efficient manner. These conventions are described in this section. This section will be updated with more specific conventions and reporting structures after CCOS measurement investigators have been commissioned and provided their suggestions and recommendations. To the greatest extent possible, CCOS field data structures, processing, validation, and delivery procedures will be consistent with those established for the long-term data base and other ARB data sets from recent air quality studies (e.g. Fujita et al., 1997).

A data manager should be appointed at least one year prior to the field program. The data manager will be responsible for establishing and maintaining computer-based data archives and managing the overall data storage and validation activities. The data manager will interact with the field manager and the program management and provide feedback to them concerning the status of unresolved data management problems and potential for resolution.

C.1 Data Specifications

CCOS data management conventions and methods build on experience from and development supported by the 1990 San Joaquin Valley Air Quality Study (SJVAQS) (Blumenthal et al., 1993), the 1995 Integrated Monitoring Study (IMS-95) (Solomon and Magliano, 1997), and the 1997 Southern California Oxidant Study (SCOS-97) (Fujita et al., 1997). The following specifications are maintained by the data manager and are available to all project participants via the internet:

- **Measurement locations:** Each measurement location is identified with a unique alphanumeric site ID accompanied by its name and address, coordinates, elevation, its primary operator, and a summary of measurements taken at the site for different monitoring periods. Appendices A-D summarizes existing and proposed field study measurement locations. Coordinates and elevations are verified by the field manager with a Global Positioning System (GPS), pressure-based altimeter, and topographical maps. Immediate surroundings are recorded with a digital camera and a video tape of the area surrounding the site is available and catalogued. The make and model of instruments used to acquire measurements at each site is recorded with its threshold, range, expected accuracy and precision.
- **Variable definitions:** Each variable is assigned a unique code that is accompanied by its definition, units, averaging time, applicable temperature and pressure adjustments, and data reporting format.
- **Data validation flags:** Flags specific to each measurement investigator are translated into a common set of validation flags that are carried with each data point. These are currently being defined by EPA for its speciation program, and this will be a starting point for CRPAQS data validation flags.

- **Data files:** Basic data files are constructed in normalized formats that have the same structure for different types of data. These files will be transparent to most users.
- **Timing conventions:** Times are expressed in Pacific Standard Time (PST), hour or minute beginning. All dates as MM/DD/YYYY (note that years are four-digit codes: 1999, 2000, 2001).
- **Missing data conventions:** Missing or invalid measurements are replaced by a –99 or a “NULL” value if the data management software permits.
- **Investigator code:** Each measurement investigator or network operator is assigned a unique two-character code that is used to identify the source of the data. Data transmittals will carry this code as part of the filename when they are loaded into a raw data sub-directory, and a separate sub-directory on the CRPAQS internet server.

C.2 Data Formats

Data acquired at set intervals are submitted as comma delimited text files via file transfer protocol to the sub-directory set up for each measurement investigator. File naming conventions are given below. Raw data file naming conventions include the investigator code, a measurement type code, and an indicator for the period of data acquisition. Formats are:

- **Continuous sequential measurements:** SSSS, MM/DD/YY, HHMM, TIME, DURATION, PARAM1, RESULT, QCFLAG, PARAM2, RESULT, QCFLAG,
- **Surface particle measurements:** SSSS, MM/DD/YY, HHHMM, TIME, DURATION, SIZE, PARAM1, RESULT, QCFLAG, PARAM2 ,RESULT, QCFLAG
- **Upper air measurements:** SSSS, MM/DD/YY, HHMM, TIME, EV_MSL, PARAM1, RESULT, QCFLAG, PARAM2, RESULT, QCFLAG

The mnemonics have the following definitions:

- **SSSS** = 3 - 4 character project site ID code
- **MM/DD/YYYY** = date specification
- **HHMM**=standard sample start time, begin hour and minutes PDT (0000-2355)
- **TIME** = actual sample start time, HHMMSS
- **DURATION** = sampling in total minutes
- **EV_MSL** = elevation in meters above mean sea level
- **PARAM** = project parameter code
- **SIZE** = 1 letter particle size category (e.g. T = PM₁₀ (0-10 µm); F = Fine(0-2.5 µm); C =Coarse (2.5-10 µm); S=sum of Fine+Coarse ~ PM₁₀; P = TSP (0-30 µm), and other particle size ranges as needed)

- **RESULT** = data value in project specified significant digits
- **QCFLAG** = 1 character project QC code indicating quality of data point

C.3 File Names

File names are of the form CDDMMYYA.PLL

- **C**=investigator code (**to be defined**)
- **DD** =data type code (**to be defined**)
- **MM** = month code (JA,FE,MR,AP,MY,JN,JL,AU,SE,OC,NO,DE)
- **YYYY** = Year code
- **A** = Averaging interval codes: (A = 3 hour; B = 6 hour; C = 12 hour; D= 24 hour ;H = 1 hour; J = jumps (every other hour); V = hourly but varies, possibly less than 24 measurements per day; I = Instantaneous (< 1 min); F=5 minute; T=10 minute; M = 15 minute; N=30 minute; P = Partial hour samples (< 60 min)
- **P** = Measurement platform code (S = surface, U = upper air)
- **LL** = Two character data validation level code (OA,OB,1A,1B,2A,2B,3A).

C.4 Validation Flags

Procedure- and investigator-specific validation flags will be maintained in a separate validation file. These must be translated into the common flags listed below. A translation table will be established as part of the database that associates each investigator flag with one of the following flags: 0=valid; 1 = estimated; 2=calibration; 3=instrument failure; 4=off-scale reading; 5 = interpolated; 6=below detection limits; 7=suspect; 8=invalid; 9=missing; a=hourly avg (45 <->60 minutes); b=hourly avg (<45 minutes); d=averaged data; e=zero mode; and f=blank sample

C.5 Data Validation Levels

Mueller (1980), Mueller et al., (1983), and Watson et al. (1983, 1989, 1995) define a three-level data validation process that should be mandatory in any environmental measurement study. Data records are designated as having passed these levels by entries in the VAL column of each data file. These levels, and the validation codes that designate them, are defined as follows:

- **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 data base 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 CRPAQS database, although they are consulted on a regular

basis to ascertain instrument functionality and to identify potential episodes prior to receipt of Level 1A data.

- **Level 1A (1A):** These data have passed several validation tests applied by the measurement investigator prior to data submission. The general features of Level 1A are: 1) removal of data values and replacement with -99 when monitoring instruments did not function within procedural tolerances; 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 1B (1B):** Pre-programmed consistency and reasonability tests are applied by the data manager prior to integration into the CRPAQS data base. 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. Others may be invalidated or flagged based on the results of the investigation. The bounds used in these tests will be determined in cooperation with measurement investigators and network operators..
- **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 2A 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.
- **Level 3 (3):** Level 3 is applied during 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 CRPAQS 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.

C.6 Internet Server

CCOS/CRPAQS data and communications and will be received and made available on the Internet server <http://sparc2.baaqmd.gov/centralca/> maintained by the Bay Area Air Quality Management District. This server currently contains project documentation and will be developed to contain project status during field monitoring and project data as it becomes available.

C.7 Directory Structure

Data and communications files are organized into several sub-directories within the CCOS/CRPAQS server. Each of these contains additional sub-directories to further organize the information. This organization will be transparent to most users who will access information through links available through browsing software. All directories, with exception of the TEMP directory and its sub-directories, have read-only privileges for most users to avoid the inadvertent erasure of information. Files may be uploaded to investigator-specific sub-directories in the TEMP directory for later placement in the appropriate read-only directory by the data manager. These CCOS directories are:

- **TEMP:** This is a temporary location where files are uploaded by project participants prior to their transfer to their designated parent directory. An e-mail message should be sent to the data manager indicating the uploaded file name and its desired directory location. The TEMP directory is also used to allow the transfer of non-archived scratch files among project participants.
- **REPORTS:** This directory contains files related to project reports, memoranda, and minutes. Sub-directories are:
 - PROGPN for the latest draft of the program and management plans
 - MEMO for project memoranda

- MINUTES for minutes of discussions
 - NOTICES for meeting notices
 - PRGREP for progress reports
 - RFP for final versions of requests for proposals
-
- **RAWDATA:** This directory contains data files in the form they are received from the data source. These files are in several different formats, cover different time periods, and do not necessarily conform to the units and variable naming conventions adopted for CRPAQS field studies. To conserve disk space, these files are usually backed up onto storage media after they have been processed; they can be re-loaded upon request to the data manager. They are located in directories specific to each data supplier, as specified in the next section.
 - **DATA:** This directory contains validated ambient measurement data in data management formats that have been converted to common units and variable names.
 - **QA:** This directory contains quality assurance results from audits, performance tests, and collocated measurements.
 - **TABLES:** This directory contains tables of processed results, including statistical summaries of data, frequency distributions, and data capture rates. These are made available via html links through the CCOS home page.
 - **FIGURES:** This directory contains figures of processed results, including time series, spatial isopleth plots, cumulative frequency plots, and scatterplot comparisons. These are made available via html links through the CCOS home page.
 - **MAPS:** This directory contains base maps of terrain, highways, population centers, political boundaries, land use, and surface characteristics in formats that are deemed useful for different analyses.
 - **UTILITIES:** This directory contains software created for the project, software available for distribution for which licenses have been obtained, and commonly used shareware. These include data conversion, format conversion, file compression, and data display programs. These are made available via html links through the CCOS home page.

Other directories and sub-directories are created as needed to organize the information produced by CCOS field studies.

C.8 Data Processing

Data are submitted by each investigator using the defined variable naming conventions and units. All values are to be at Level 1A when submitted. These are passed through the Level 1B tests described above, and discrepancies are resolved with the measurement investigator and corrected prior to designation as Level 1B. Data are added to the master data files as they are received.

Measurements from other ongoing networks described in Section 4 are acquired in formats and units specific to those networks. Data processing functions are specific to these networks and have been established during the compilation of the multi-year database for central California. First, these files are manually edited, when needed, because they sometimes contain minor variations in format that confound the data conversion programs. After editing, a network-specific data conversion program reads the data and converts it to an xBase file format with a single record for each measurement. Variable names in these intermediate files differ from those in CCOS files by designating the unit used for each measurement in the specified network. For example, field name “TA__F” indicates ambient temperature in degrees Fahrenheit as found in the NWS database rather than the “TA” field name for ambient temperature in degrees Celsius that is used in the CCOS data files.

The next step converts measurement units to the CCOS common units. Conversion factors are accessed from a file that maps one unit into another based on the specification of the input and output variable names. In addition to changing units, the conversion program maps times, dates, missing values, and validation flags into the CCOS conventions described above.

A data validation log is kept to document all changes made to the data files, including changes in the data validation level. This includes a record of the data changed, the reason for the change, and the date of the change. All data as originally submitted to the database are retained.