9. Centralized Data Management and Validation

Overview

EPA/EMSL in Las Vegas will be the data managers for Project MOHAVE. Information will be obtained from the following sources:

SOURCE	DATA TYPE
NOAA/WPL	Surface and Upper Air Meteorology
Brookhaven National Laboratory	Tracer concentrations
UC-Davis	Aerosol and SO_2
Air Resource Specialists	Optical and Surface Meteorology
EPA-RTP/AREAL	Aerosol
National Meteorological Center	Surface and Upper Air Meteorology
Colorado State University	Meteorological Modeling
CAPITA- Washington University	Monte Carlo Modeling

The data to be collected are described in more detail in Sections 3-7. A data management and validation plan will be developed by the data management coordination committee. A sketch of the expected elements of the plan to be developed is presented in the remainder of this section.

Two levels of validation (Levels 1 and 2) will be systematically applied. Level 1 (univariate) validation involves checking the data for outliers, rates of change, proper indication of time and location of data, etc. In Level 2 (multivariate) validation, consistencies among variables and the appropriateness of spatial and temporal patterns are investigated. For example, the light scattering (b_{scat}) measured by a nephelometer should be less than the total extinction (b_{ext}) obtained by a transmissometer. Level 3 validation occurs during the data analysis. If data inconsistencies are found, the documentation regarding the questionable observation is examined for correctable errors (e.g. transcription errors). Uncorrectable, suspect data are flagged, but not removed from the data set. Data known to be incorrect and not recoverable are removed from the data set.

Each group responsible for collecting data will perform at least Level 1 validation. UC-Davis will do a partial Level 2 validation of the aerosol data. The data managers at EPA/EMSL Las Vegas are responsible for the Level 2

validation. Systematic procedures and protocol for the Level 2 validation will be developed and fully documented prior to releasing the data. Level 1 protocols utilized by organizations responsible for each data subset will also be documented. A computerized listing of the data will be prepared. Level 2 data will be distributed to data analysts and other interested parties. At the end of the study, all data will be assembled and documented. A brief discussion of validation conducted by some of the participants appears in the following subsections.

Aerosol Sampling (UC-Davis)

A number of the measured or derived parameters are interrelated. This allows data intercomparisons as a method to evaluate system performance and check for outliers. The intercomparisons made are listed below:

(1) Fine sulfur vs. fine sulfate
(2) Fine sulfur vs. PM-10 sulfur
(3) Fine hydrogen vs. fine mass
(4) PM-10 hydrogen vs. PM-10 mass
(5) Sum of fine components vs. fine mass
(6) Sum of PM-10 components vs. PM-10 mass
(7) Elemental carbon vs. optical absorption
(8) Organic carbon vs. nonsulfate hydrogen
(9) Fine mass vs. extinction
(10) PM-10 mass vs. extinction
(11) Fine mass components vs. extinction
(12) PM-10 mass components vs. extinction

Details of the quality assurance and data validation are given in Pitchford and Joseph (1990).

Transmissometer Data

The transmissometer data is subjected to three levels of validation. In the first level, validity codes reflecting transmissometer instrument operation are added to the raw transmissometer data files. In the second level, data and validity codes are checked for inconsistencies using a screening program. The b_{ext} data are adjusted for lamp drift of 2% per 500 hours of lamp-on time. Validity codes are added to all data. The third level, consists of 2 steps,

(1) Calculation of uncertainty values for all data; and

(2) Identification of b_{ext} values affected by weather.

Validity codes for b_{ext} include:

0 = valid
1 = Invalid: Site operator error
2 = Invalid: System malfunctioned or removed
3 = Valid: Data reduced from alternate logger
4 = Weather: Relative Humidity > 90%
5 = b_{ext} > maximum threshold
6 = Δb_{ext} > delta threshold
7 = b_{ext} uncertainty > threshold
8 = Missing: Data acquisition error
9 = Invalid: b_{ext} below Rayleigh
A = Invalid: misalignment
L = Invalid: Defective lamp
S = Invalid: Suspect data
W = Invalid: Unclean optics

Radar wind profilers and RASS

Real-time processing consists of a Doppler spectra peak picking routine which searches for spectra peaks beginning from the highest level of good signals to the lowest gate. As the routine searches for peaks in a downward direction, it requires consistency from gate-to-gate. If a peak shifts beyond a given threshold between gates, that peak is rejected. To help eliminate ground clutter, the algorithm also rejects peaks near zero velocity if a secondary peak away from zero is available. After the peaks, or first moments, for each individual radial are chosen in this way, a consensus averaging is performed. This technique requires at least 50% of the points on each gate of each radial for a 55 minute period (5 minutes for RASS temperature) to fall within a bin of 2 m/s in width before the individual points in the bin are averaged. If less than 50% of the points fall within the bin, the radial component is flagged as bad and is not available for that period. A similar technique is used for the RASS derived temperatures with a bin threshold of 1 °C.

The normal post-processing quality assurance procedures consist of applying a time/height editor, normally referred to as the Weber/Wuertz editor (Weber and Wuertz, 1989), to each 24 hour period of one hour averaged profiler wind data or 5 minute averaged temperature data (one 5 minute average provided each hour). The editor assesses the neighborhood of each point for consistency in both speed and direction (or temperature), allowing for a larger tolerance for direction differences at lighter wind speeds. The tolerances are adjustable and depend on the prevailing meteorology during a particular experiment. The neighborhood size is also adjustable, but usually the eight adjacent points are chosen, if available. This editor has proven to be very powerful in eliminating outlier points. The results of this processing provide the Level 1 data.

NOAA/WPL is experimenting with applying a more sophisticated form of this editor on the radial moments before performing an hourly average. The test data are from the 1990 San Joaquin Valley Air Quality Study. The technique requires considerable processing. The decision to use this technique for Project MOHAVE depends on the quality of the data, which in turn depends on site characteristics.

The post-processed, Level 1 data will be compared with optically tracked rawinsonde (airsonde) wind and temperature profiles measured at each location. Several rawinsonde profiles will be available at each of the wind profiler locations representing different stability and meteorological conditions at each site.