Data Quality of Quality Measurement Experiments

S. Bottone and S. Moore
Mission Research Corporation
Santa Barbara, California

Introduction

Quality Measurement Experiments (QME) are a special class of Value-Added Products (VAP). QMEs add value to Atmospheric Radiation Measurement (ARM) Program datastreams by providing for continuous assessment of the quality of incoming data based on internal consistency checks, comparisons between independent similar measurements, or comparisons between measurements and modeled results. Like any datastream, QME datastreams need to be checked for data quality. For each QME, we analyze a representative sample of files from the ARM data archive to determine 'typical' values of the QME variables. We then design outlier tests, specific to each variable, to be applied to future data. If a measurement is determined to be inconsistent with previous measurements (i.e., is an outlier), it is flagged as a possible data quality problem. For large datasets, measurements may be grouped and given one quality flag. For each QME, quality information is presented in color-coded tables, available on the Data Quality Health and Status (DQ HandS) website and organized by date, which provide the data quality of relevant variables for the QME. We present data quality analyses of two QME's: QMEMWRPROF – which compares retrieved water vapor and temperature profiles from the VAP *mwrprof* with radiosonde profiles; and QMEMWRCOL – which compares ARM microwave radiometers (MWRs) against output from an Instrument Performance Model (IPM).

Description of QME's

We performed a data quality analysis on two QMEs, which includes a statistical analysis of historical data and interactive data language (IDL) programs that perform quality tests that flag data with possible data quality problems and generate corresponding diagnostic plots:

(From http://www.arm.gov/docs/research/vap_homepage/details/*)*

QMEMWRPROF: This QME compares the retrieved water vapor and temperature profiles from the VAP *mwrprof* with radiosonde profiles. The experiment is designed to help evaluate the skill of the retrieval algorithm used in the VAP *mwrprof*. In addition to the retrieval algorithm, this QME provides a check on the measurements made by the ground-based sensors used as input to the retrieval algorithm. This QME also helps to assess the ability of the radiosondes to measure the moisture and temperature in the atmosphere.

Output platform: sgpqmemwrprofC1.c1 – water vapor and temperature residuals and statistics between the retrieved (*mwrprof*) values and observed (radiosonde) values.

QMEMWRCOL: The first QME implemented in the ARM Experiment Center, this VAP compares the ARM MWRs against the output from an IPM, which is based upon the Leibe (1987) microwave absorption model. The thermodynamic profile from the radiosonde is used to drive the model, which outputs brightness temperatures at the same frequencies as the radiometer. The results from this QME are used to evaluate the MWR and radiosondes, update the tuning functions used in the MWR, and check the MWR's calibration.

Output platforms: [*site*]qmemwrcol[*location*].c1 – model calculated brightness temperatures, residuals of brightness temperatures and total precipitable water vapor.

Approach

For each QME we analyze a representative sample of ARM output files from the data archive to determine 'typical' values of the output variables. We then design outlier tests, specific to each variable, to be applied to future data. If a measurement is determined to be inconsistent with previous measurements (i.e., is an outlier), it is flagged as a possible data quality problem. Results are made available in color-coded tables, along with diagnostic plots, on the DQ HandS website.

Datasets

We have fetched over 6000 NetCDF output files from the ARM data archive. Table 1 gives the test sites and other characteristics of these datasets.

Table 1. Datasets

Test Site	Platform	Start Date	End Date	Number of Files	Number of Times
SGP	sgpqmemwrprofC1.c1	19950420	20001231	1112	7998
SGP	sgpqmemwrcolC1.c1	19960111	20010709	1486	6341
SGP	sgpqmemwrcolB1.c1	19960101	20001008	637	1822
SGP	sgpqmemwrcolB4.c1	19960101	20001127	621	1755
SGP	sgpqmemwrcolB5.c1	19960101	20001008	555	1535
SGP	sgpqmemwrcolB6.c1	19960101	20001008	640	1699
NSA	nsaqmemwrcolC1.c1	19980606	20010724	643	667
TWP	twpqmemwrcolC2.c1	19981120	20010331	766	1496

Analysis of QMEMWRPROF

Each NetCDF file in the output platform, sgpqmemwrprofC1.c1, contains 20 variables, measured (or retrieved) at several times in a given day. We performed a data quality analysis on the four most relevant values:

temp_resid Retrieved absolute temperature minus sonde absolute temperature

rho_resid	Retrieved water vapor density minus sonde water vapor density
temp_RMSE	Root mean square error of temperature residuals
rho_RMSE	Root mean square error of water vapor density residuals

Nearly 8000 temp_resid and rho_resid profiles over a 6-year period, which consist of values at 49 heights ranging from 0 to 12 km at 0.25 km steps, were analyzed to estimate mean values and standard deviations at each height. Histograms at each height were examined and it was determined that values that are more than four standard deviations from the mean can be considered outliers. Histograms of temp_RMSE and rho_RMSE were also examined and outliers were chosen to be those values of temp_RMSE that exceed 20°C and rho_RMSE that exceed 3 g/m^3. Histograms of temp_RMSE and rho_RMSE over a 6-year period are shown in Figure 1.

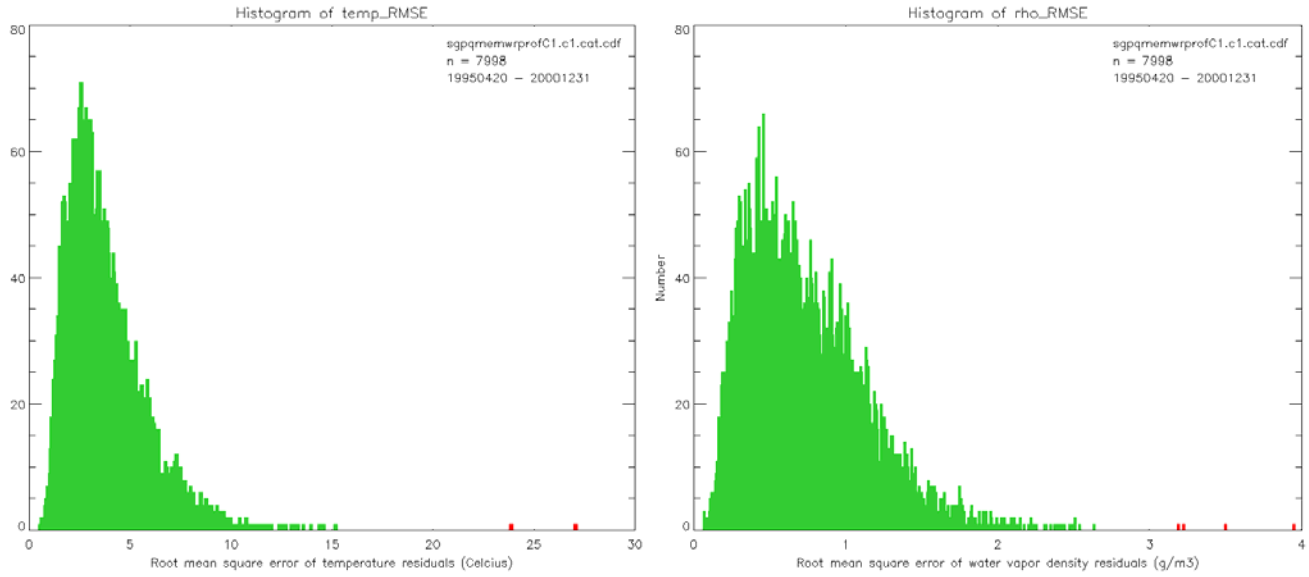

Figure 1. Histogram of temp_RMSE and histogram of rho_RMSE. Outliers are marked in red.

Quality tests for QMEMWRPROF can be summarized as follows:

mean outlier	If temp_resid or rho_resid at a given height deviates more than four standard deviations from its six-year mean then it will be flagged
RMSE	If temp_RMSE for a profile exceeds 20°C then the entire temp_resid profile is flagged If rho_RMSE for a profile exceeds 3 g/m^3 then the entire rho_resid profile is flagged
missing	If data value is missing then it is flagged.

We have developed and implemented an IDL program, called *dq_qmemwrprof*, which constructs files necessary to produce tables to be incorporated into the DQ HandS website. In addition, diagnostic plots, to be displayed on the web, are also produced by the program. The input to the program is any sgpqmemwrprofC1.c1 NetCDF file. For each time in the input file, two output files are produced, one of the form *platform.date.time.dat*, which contains the information necessary to create the table, and the

other of the form *platform.date.time.gif*, which contains a diagnostic plot for that time. Figure 2 shows a typical diagnostic plot for a case where quality flags indicate that temp_RMSE has exceeded 20°C and values of temp_resid are greater than 4 standard deviations from the mean at several heights.

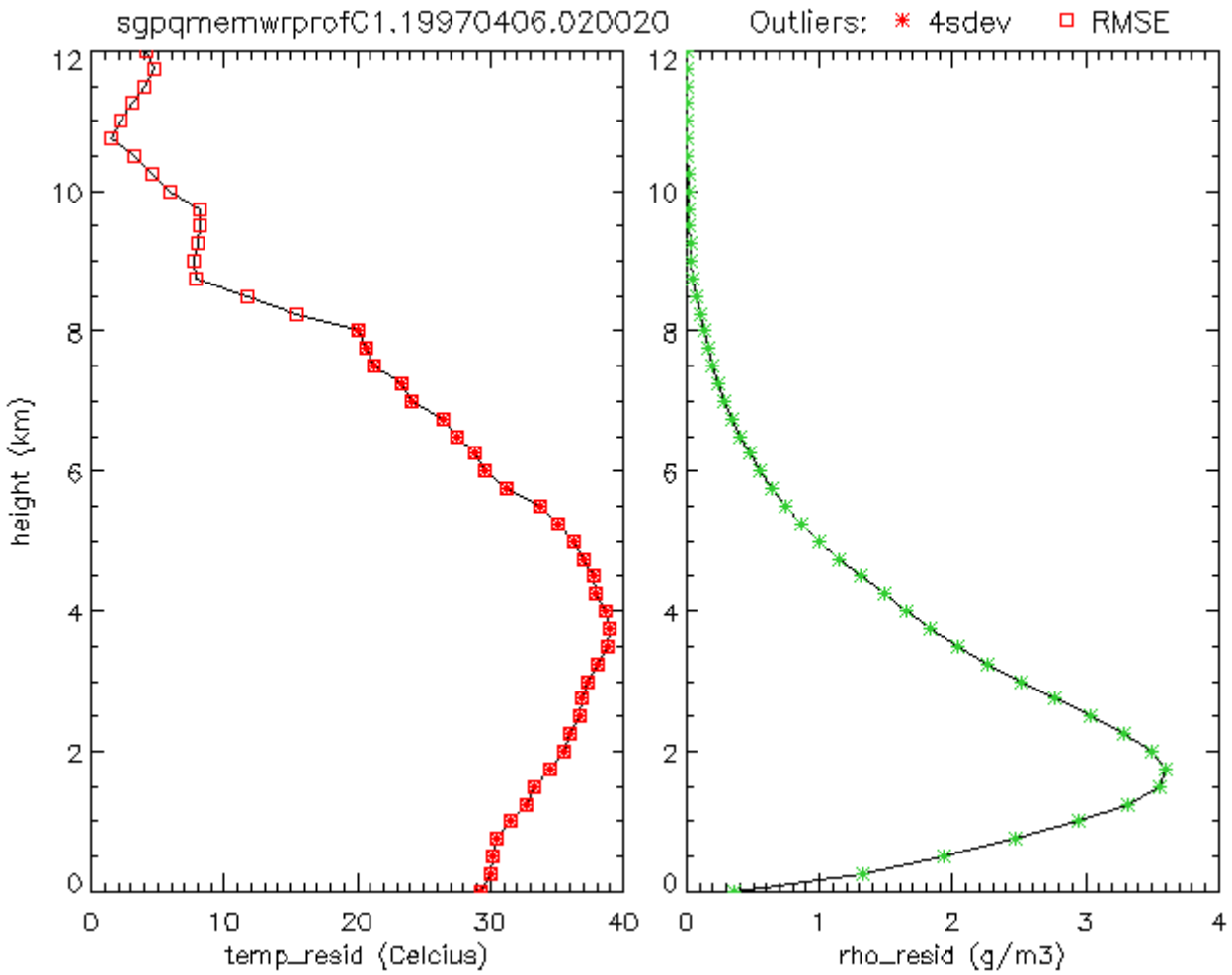

Figure 2. Sample diagnostic plot from the IDL program, *dq_qmemwrprof*, showing plots of height (km) vs temp_resid and rho_resid. The temp_resid profile has exceeded the RMSE threshold of 20°C, while the rho_resid profile passes all data quality tests.

Analysis of QMEMWRCOL

Each NetCDF file in the output platform, [*site*]qmemwrcol[*location*].c1, contains 37 variables, measured (or retrieved) at several times in a given day. We performed a data quality analysis on the four most relevant values:

mean_vap_mwr:	Ensemble average for MWR vapor in window centered about balloon release
mean_tbsky23_mwr:	Ensemble average for MWR 23.8 GHz sky brightness temperature in window centered about balloon release

mean_tbsky31_mwr: Ensemble average for MWR 31.4 GHz sky brightness temperature in window centered about balloon release

qc_flag_mwr_set: Flag that is set to 1 if any Mentor QC flag is set for the mwr for the fields tbsky23, tbsky31, vap, liq, and ir_temp; 0 if none of these flags are set

Over 6000 [*site*]qmemwrcol[*location*].c1 files were fetched from the ARM data archive and analyzed. These files contain ensemble averages of water vapor content and sky brightness temperatures at 23.8 and 31.4 GHz at an average of about 4 times per day. Each ensemble average represents an average over forty minutes of MWR data around times corresponding to a radiosonde release. The corresponding radiosonde release provides input for the IPM model, which outputs to the NetCDF files water vapor and brightness temperatures, which can be compared to the corresponding MWR averages. The values are labeled:

integ_vap_sonde: Integrated vapor column from sonde using MWR Instrument Performance Model (IPM)

model_tbsky23: MWR IPM output for 23.8 GHz sky brightness temperature using sonde T,P,RH

model_tbsky31: MWR IPM output for 31.4 GHz sky brightness temperature using sonde T,P,RH

Quality tests for QMEMWRCOL can be summarized as follows:

model comparison: If the normalized difference of mean_vap_mwr, mean_tbsky23_mwr, or mean_tbsky31_mwr with its corresponding model value deviate more than three standard deviations from its five-year mean then that variable will be flagged

minmax: If mean_vap_mwr, mean_tbsky23_mwr, or mean_tbsky31_mwr is less than the minimum or greater than the maximum given in table 2 then that value is flagged

qc_flag: If qc_flag_mwr_set equals one then a separate flag is set

missing: If data value is missing then it is flagged

The model comparison test, which compares the normalized difference of an MWR measurement with a corresponding model prediction, is as follows: flag test value x_1 as an outlier if

$$\left| \frac{x_1 - x_2}{\sqrt{s_1^2 + s_2^2}} - \mu \right| > 3$$

where,

x_1 = test value (mean_vap_mwr, mean_tbsky23_mwr, or mean_tbsky31_mwr)

x_2 = corresponding model value

s_1 = standard deviation of test value

s_2 = standard deviation of model value
μ = five-year mean of normalized difference of x_1 and x_2

For the minmax test, the minimum and maximum values are given in Table 2.

Table 2. Minimum and maximum values of minmax test.

	SGP	NSA	TWP
min_vap (cm)	0	0	0
max_vap (cm)	6	3	7
min_tbsky23 (K)	2.73	2.73	2.73
max_tbsky23 (K)	80	60	100
min_tbsky31 (K)	2.73	2.73	2.73
max_tbsky31 (K)	40	30	50

We developed and implemented an IDL program, called *dq_qmemwrcol*, which constructs files necessary to produce tables to be incorporated into the DQ HandS website. In addition, diagnostic plots, to be displayed on the web, are also produced by the program. The input to the program is any [*site*]qmemwrcol[*location*].c1 NetCDF file, which typically consists of data at several times for a given date. For each input file, two output files are produced, one of the form *platform.date.dat*, which contains the information necessary to create the table, and the other of the form *platform.date.gif*, which contains a diagnostic plot for that date. Figure 3 shows a typical diagnostic plot. The figure contains plots of mean_vap_mwr, mean_tbsky23_mwr, and mean_tbsky31_mwr vs the corresponding model predictions for each time in the file. In this case there are some data values that are outliers and/or qc_flag_mwr_set is equal to one. In addition, if data is missing in the file, the plot will give the number of missing times.

Acknowledgment

This research was supported by the Biological and Environmental Research Program (BER), U.S. Department of Energy, Grant No. DE-FG03-97ER62468.

Corresponding Author

Steven Bottone, bottone@mrcsb.com, (805) 963-8761, ext 319, http://arm.mrcsb.com

Reference

Leibe, H. J., and D. H. Layton, 1987: Millimeter wave properties of the atmosphere: Laboratory studies and propagation modeling. NTIAREP., 87-24, p. 74.

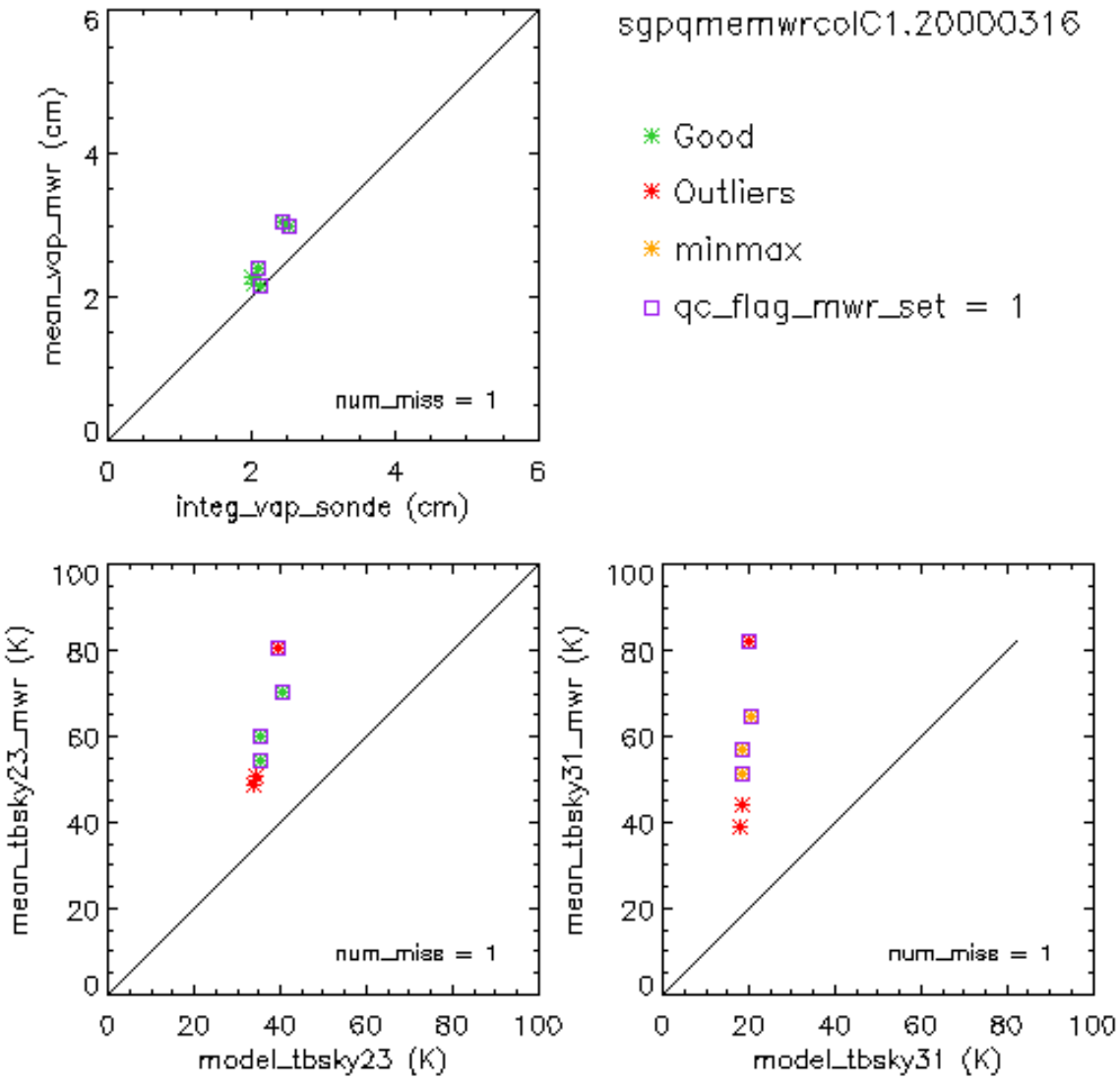

Figure 3. Sample diagnostic plots from the IDL program, *dq_qmemwrcol*, showing plots of mean_vap_mwr, mean_tbsky23_mwr and mean_tbsky31_mwr vs model predictions. Some data fails outlier test or minmax test and some data has qc_flag_mwr_set = 1. There is also one missing data value.