NARSTO MODEL INTER-COMPARISON (NMI) STUDY

DRAFT

DATABASE DOCUMENTATION
STI-900560-2111-DD

By:
Neil J. M. Wheeler
Jason Roney
Sonoma Technology, Inc.
1360 Redwood Way, Suite C
Petaluma, CA 94954-1169

Prepared for:
Texas Natural Resource Conservation Commission
P.O. Box 13087
Austin, TX 78711-3087

August 23, 2001

1. INTRODUCTION

A major study comparing regional air quality modeling systems used for regulatory purposes in the United States and Canada is being conducted under the auspices of the North American Research Strategy for Tropospheric Ozone (NARSTO). The opportunity for such a study arose when, stimulated by the need to meet regulatory milestones for attainment of national standards and objectives for ozone in the two countries, many public and private groups modeled time periods in July 1995 in overlapping domains covering most of eastern North America. These modeling activities included the application of several different meteorological and air quality models, and used data from the NARSTO-Northeast (NARSTO-NE) and Southern Oxidant Study (SOS) field studies. The study began taking shape at a workshop for interested parties in May 1998. Work groups focusing on emissions, meteorological and air quality modeling were formed and charged with developing comparison and evaluation protocols for models in their respective areas.

The study was conceived to take place in two phases. In the first phase, the models were to be compared in their "native" mode, i.e., using the same input files and model configurations that each group used in their individual assessments, based on modeling a 12- to 14-day period in July 1995. The planners believed that, with a minimal amount of extra effort, existing model output files could readily be used for the comparison and valuable information would be gained on the relative performance of these disparate models, each of which was used in exercises having significant policy implications.

However, based on the realization that such an approach would provide insufficient information to diagnose why models might yield different results, the second phase would approach the problem in a more comprehensive manner. It would also recognize the anticipated widespread desire to compare models with an aerosol simulation capability. Thus, in Phase 2, regional tropospheric aerosol models would be compared using harmonized model inputs. "Harmonization" means that models would be exercised, for example, on identical domains, employing the same horizontal gridding, topography, land use, vegetation distributions, and emissions inputs. Meteorological inputs would be as similar as possible, consistent with maintaining mass conservation.

Phase 1 of the study was partially completed in August 2000 and a workshop was held to discuss the results up to that point and plan the future activities. **Table 1** shows the models that were included in the inter-comparison. A total of 11 base-case simulations were developed with five of the six participants providing simulations at two different grid resolutions.

Because aerosol modeling was still in its infancy, Phase 2 of the study could not be launched at this stage. Instead, a Phase 1B was planned to be completed by July 2001. In Phase 1B, simulated and observed values will be compared for base-case runs, model-to-model comparisons of meteorology and air quality will be made, and emissions sensitivity simulations will be run and analyzed. Before these tasks can be carried out, observational data must be obtained and results extracted from the model output files. Sonoma Technology, Inc. (STI) has acquired, checked, standardized, and documented the observational data needed. STI has also processed the model output provided by the study's participants to extract the data required to

perform model-to-model and model-to-observation comparisons. STI is providing data sets containing observations and the extracted model output. This document describes the methods used to develop these data sets and the electronic formats used in creating them. Both TNRCC and the study's managing organization, the Electric Power Research Institute (EPRI) will receive these data with this documentation.

Table 1. Models used in inter-comparison study.

Participant/Sponsor	Meteorological Model	Air Quality Model	Grid-cell Size, km	Period Simulated
Meteorological Service of Canada	MC2	CHRONOS	40 and 10	7-18 July, 1995
U.S. EPA	MM5	CMAQ	36 and 12	7-18 July, 1995
New York Department of Environmental Conservation	RAMS3b	UAM-V	36 and 12	7-18 July, 1995
North Carolina Supercomputing Center	MM5	MAQSIP	36	7-18 July, 1995
Environ/Coordinating Research Council	MM5	CAMx	36 and 12	7-15 July, 1995
ICF Consulting/Southern Co.	MM5	UAM-V	36 and 12	7-15 July, 1995

2. DATABASE DOCUMENTATION

As a part of the NARSTO Model Inter-comparison (NMI) study, STI acquired and processed both observational and model output data to be used in the study. This section describes the data acquired, how the data were processed, and the formats used in the NMI database.

2.1 DATA SOURCES

2.1.1 Observational Data

Observational data were obtained from four sources:

1) Electric Power Research Institute (EPRI)

2) National Aeronautics and Space Administration (NASA)

3) NARSTO-NE Carbonyl

4) Southern Oxidants Study (SOS).

The NASA database includes observations from the NARSTO-NE database. The types of data provided in these databases are hourly information on O_3, NO, NO_2, NO_y, and NO_x for the period of July 1995. In addition, 3-hr and some 6-hr average HCHO (formaldehyde) data were obtained from the NARSTO-NE archive, which is maintained by STI.

2.1.2 Model Output Data

Model output was provided by EPRI on Exabyte 8-mm tapes. A summary of the model output provided is shown in **Table 2**. Only base-case simulations were provided.

Table 2. Model data provided by EPRI.

Participant/Sponsor	Air Quality Model	Course Grid Average	Fine Grid Average	Instantaneous
NY Dept. of Environmental Conservation (NYDEC)	UAM-V		X[1]	
North Carolina Supercomputing Center	MAQSIP			X
Environ/Coordinating Research Council	CAMx	X	X	
ICF Consulting/Southern Co.	UAM-V	X	X	

[1] A hybrid course-fine (*.cf) grid average file was provided from which only the fine grid average data can be extracted as the course grid data is interpolated to the fine grid.

2.2 PROCESSING

2.2.1 Observations

Processing of the observational data sets was designed to extract both the species of interest and the dates of interest. For this study, the species of interest were O_3, NO_y, NO_x, and HCHO. Where possible, NO_x was constructed as the addition of NO and NO_2. The dates of interest extracted included July 4 to 19, 1995 (if available), which contain the dates simulated in the models.

The observational data was put in common format for the previously mentioned data sets in four subsets, SOS.dat, NASA.dat, NARSTO.dat, and EPRI.dat. The initial processing required multiple FORTRAN programs to read the various formats as each set varied slightly in format. Common format differences included species, date format, and time zones. Where necessary, programs were written to account for time zone changes, and all times were written to Eastern Standard Time (EST). The common format included two header lines and then data of interest as shown below:

```
Data from SOS, times are EST
QUALCODE,STNID,STATE,SOURCE,DATE,TIME(EST),DUR,O3,O3QC,NOX,NOQC,NOY,NOYQC,HCHO,HCHOQC
1,'010070001','AL','SO','07/04/1995', 0, 60,  14.308,0,-999.000,9,-999.000,9,-999.000,9
```

These data sets were then merged with another program to a final format that is discussed in Section 2.3.

2.2.2 Model Results

The model output data were extracted after modifying the provided FORTRAN programs listed in **Table 3**. Typical modifications to the programs included providing the ability to read multiple species, and multiple dates, and extracting those sites only relevant to the regions of interest. In addition, extraneous parts of the program were eliminated, specifically, those involving simultaneous extraction of the observational data.

Table 3. Programs provided for extraction.

Fortran Program Provided	Model Extraction	Grid-Type	Output
cc_outcon_camx.f	CAMx	Course, 36 km	1-hr average concentrations
ff_outcon_camx.f	CAMx	Fine, 12 km	1-hr average concentrations
cc_outcon_sai.f	UAM-V	Course, 36 km	1-hr average concentrations
ff_outcon_sai.f	UAM-V	Fine, 12 km	1-hr average concentrations
mcnc5.f	MAQSIP	Coarse, 36 km	1-hr average concentrations

Using the modified programs, each species of interest was written to a separate file. Within each of these files, the date, hour, relevant observational site and corresponding grid-cell

concentrations for the model data are present. A site identification list was provided for this extraction and is discussed later. The 10 main species extracted are the following: O_3, NO, NO_2, NO_3, N_2O_5, N_xO_y, HNO_3, HONO, PAN, and HCHO. The Environ and the ICF model output each contained all the previous species except NO_3 and N_2O_5, which were incorporated in the term N_xO_y. On the other hand, the MCNC model output contained all the previously mentioned species including NO_3 and N_2O_5, but did not have N_xO_y. For the NYDEC model output, a modified version of the "cc_outcon_sai.f" program was used to extract the fine grid average information from the course-fine (*.cf) output. In addition, the output species differed for this model output; O_3, NO, NO_2, NO_y, HNO_3, PAN, and HCHO were extracted. Typical extracted model data using the modified programs is shown below.

```
The species extracted:   O3
siteid,date, Time,Conc.(ppbV),region,state
010270001 950707  0   33.36 USSE AL
010270001 950707  1   31.60 USSE AL
010270001 950707  2   30.36 USSE AL
010270001 950707  3   29.14 USSE AL
```

The above information for each model is in a "raw" format as extracted with the modified programs. The next step was to put each species in a common format. A program was written to write each species in a similar format to the observational data. This program accounted for the differences in date formats between models as well. A typical format showing the first three lines (2 header lines and 1 data line) is shown below for the species PAN. Each of these files contained only one species. A merge program was then used to create complete sets of species for the fine and course grid for each model. This final format is discussed in Section 2.3.

```
Extracted Data from the Model
QUALCODE,STNID,STATE,SOURCE,DATE,TIME(EST),DUR,O3,O3QC,NO,NOQC,NO2,NO2QC,NO3,NO3QC,NXOY,NXOYQC,N2
O5,N2O5QC,HNO3,HNO3QC,HONO,HONOQC,HCHO,HCHOQC,PAN,PANQC
0,'010270001','AL','SAIF ','07/07/1995', 0,60,-999.000,9,-999.000,9,-999.000,9,-999.000,9,-
999.000,9,-999.000,9,-999.000,9,-999.000,9,-999.000,9,   0.040,2
```

Again, the one exception to the above formatting is the NYDEC data, which has a similar format but instead of "NXOY" and "NXOYQC" in the header, "NOY" and "NOYQC" appear as shown below.

```
Extracted Data from the Model
QUALCODE,STNID,STATE,SOURCE,DATE,TIME(EST),DUR,O3,O3QC,NO,NOQC,NO2,NO2QC,NO3,NO3QC,NOY,NOYQC,N2O5
,N2O5QC,HNO3,HNO3QC,HONO,HONOQC,HCHO,HCHOQC,PAN,PANQC
0,'010270001','AL','NYDEF','07/07/1995', 0,60,-999.000,9,-999.000,9,-999.000,9,-999.000,9,-
999.000,9,-999.000,9,-999.000,9,-999.000,9,-999.000,9,   1.040,2
```

2.3 FORMATS

2.3.1 Site File

A site file has been created for a unique set of stations corresponding to the final observational data set. Stations with identical locations but different names were checked to see if each had the same corresponding data. Sites with unique data were listed separately. Sites

with the same data required that a single ID be chosen—in that case, an AIRS ID was given precedence. Sites without AIRS IDs were given "AIRS-like" IDs. The final site file list ("sitelist.txt") is a text file in the format shown in **Table 4**, and an example extracted from the list is shown below.

```
010070001     Centerville            AL    32.8900  -87.2300
```

Table 4. Site file format

Variable	Columns	Format
Site ID	1-9	Character (A9)
Name	15-39	Character (A25)
State	43-44	Character (A2)
Latitude	49-55	Real (F12.4)
Longitude	57-65	Real (F12.4)

2.3.2 Observations

The final observational data set included two files, one with the hourly average O_3, NO_x, and NO_y, and the other with 3-hr and 6-hr HCHO. In merging the observational data, only unique sets of data were kept. All site IDs in the observational data set correspond to those in the final site ID list. A precedence of writing was given in the merge process such that the unique observational data with the most confidence was always kept. The precedence for overwriting from least to most confident was SOS, NASA, and EPRI. Examples of the formats for the two files are shown below.

```
Air Quality Data for NMI; Times are start hour EST
QCLEVEL,STNID,STATE,SOURCE,DATE,TIME,INTERVAL,O3,O3QC,NOX,NOXQC,NOY,NOYQC
1,'010070001','AL','NMI','07/04/1995', 1, 60,  13.00,0,-999.00,9,-999.00,9

Carbonyl Data from NARSTO-NE; times are start hour EST
QCLEVEL,STNID,STATE,SOURCE,DATE,TIME(EST),DUR(MIN),HCHO,HCHOQC
1,'090031003','CT','NARSTONE','07/05/1995', 2,180,   4.41,0
1,'090031003','CT','NARSTONE','07/05/1995',14,180,   4.84,0
```

As shown, there are two header lines; the first line is a descriptive line describing the source of the air quality data; and the second line describes the information obtained in the file. The data is listed below the two header lines, and there are 13 fields, comma separated, on each line for the "NMI" data as shown and described in **Table 5**. For the "NARSTONE" data set there are 9 fields as shown and described in **Table 6**.

Table 5. Observation file format for "NMI"

Field	Variable	Description
1	Quality Control Level	See description of the QC code in Table 9.
2	Site ID	An AIRS or AIRS-like site ID in single quotes

3	State	State in which the monitor operates in single quotes
4	Source	Source of the data in single quotes (‘NMI’)
5	Date	Date in single quotes (‘mm/dd/yy’)
6	Time (Eastern Standard)	Time from 0 to 23 in EST
7	Duration of the Observation	Duration of the observation in minutes
8	Ozone Observation (ppb)	Ozone observations
9	Ozone Quality Check Code	See description of the QC code in Table 10.
10	NO_x Observation (ppb)	NO_x observations
11	NO_x Quality Check Code	See description of the QC code in Table 10.
12	NO_y Observation (ppb)	NO_y observations
13	NO_y Quality Check Code	See description of the QC code in Table 10.

Table 6. Observation file format for “NARSTONE”

Field	Variable	Description
1	Quality Control Level	See description of the QC code in Table 9.
2	Site ID	An AIRS or AIRS-like site ID in single quotes
3	State	State in which the monitor operates in single quotes
4	Source	Source of the data in single quotes (‘NARSTONE’)
5	Date	Date in single quotes (‘mm/dd/yy’)
6	Time (Eastern Standard)	Time from 0 to 23 in EST
7	Duration of the Observation	Duration of the observation in minutes
8	HCHO Observation (ppb)	HCHO observations
9	HCHO Quality Check Code	See description of the QC code in Table 10.

2.3.3 Model Results

A merge program was used to create complete sets of species for each fine and course grid for each model. In these sets, a complete site list was written and those sites without corresponding model data in the modeling domain or in one of the four specified regions were written as missing. This final merge provided six total extracted sets of model data. A typical format type below shows the first three lines in the file.

```
Model Data Extraction for MCNC; Times are start hour EST
QCLEVEL,STNID,STATE,SOURCE,DATE,TIME,INTERVAL,O3,O3QC,NO,NOQC,NO2,NO2QC,NO3,NO3QC,NXOY,NXOYQC,N2O
5,N2O5QC,HNO3,HNO3QC,HONO,HONOQC,HCHO,HCHOQC,PAN,PANQC
0,'010270001','AL','MCNC ','07/05/1995', 0, 60,  15.67,2,   0.00,2,   3.00,2,   0.00,2,-999.00,9,
0.00,2,   0.02,2,   0.00,2,   4.36,2,   0.64,2
```

As shown, there are two header lines; the first line is a descriptive line describing the type or source of the model data, and the second line describes the information obtained in the file. The data are listed below the two header lines, and there are 27 fields, comma separated, on each line for the data as shown and described in **Table 7**. For the NYDEC data set, there are 27 fields as shown and described in Table 7 as well, but in field 16 and 17, "NO_y" replaces "N_xO_y".

Table 7. Extracted model data format.

Field	Variable	Description
1	Quality Control Level	See description of the QC code in Table 9.
2	Site ID	Site ID matching a cell in the model domain in single quotes
3	State	State in which the model cell is found in single quotes
4	Source	Source of the model data in single quotes
5	Date	Date of the simulation in single quotes ('mm/dd/yy')
6	Time (Eastern Standard)	Time from 0 to 23 in EST
7	Duration	Averaging time for model data
8	Ozone Concentration (ppb)	Ozone extracted model data
9	Ozone Quality Check Code	See description of the QC code in Table 10.
10	NO Concentration (ppb)	NO extracted model data
11	NO Quality Check Code	See description of the QC code in Table 10.
12	NO_2 Concentration (ppb)	NO_2 extracted model data
13	NO_2 Quality Check Code	See description of the QC code in Table 10.
14	NO_3 Concentration (ppb)	NO_3 extracted model data
15	NO_3 Quality Check Code	See description of the QC code in Table 10.
16	N_xO_y Concentration (ppb)	N_xO_y extracted model data
17	N_xO_y Quality Check Code	See description of the QC code in Table 10.
18	N_2O_5 Concentration (ppb)	N_2O_5 extracted model data
19	N_2O_5 Quality Check Code	See description of the QC code in Table 10.
20	HNO_3 Concentration (ppb)	HNO_3 extracted model data
21	HNO_3 Concentration (ppb)	See description of the QC code in Table 10.
22	HONO Concentration (ppb)	HONO extracted model data
23	HONO Quality Check Code	See description of the QC code in Table 10.
24	HCHO Concentration (ppb)	HCHO extracted model data
25	HCHO Quality Check Code	See description of the QC code in Table 10.
26	PAN Concentration (ppb)	PAN extracted model data
27	PAN Quality Check Code	See description of the QC code in Table 10.

The source of the model data was designated in each of the files according to the model type and/or the provider of the data. **Table 8** shows the designations.

Table 8. Model data source designations.

Participant/Sponsor	Air Quality Model	Course Grid Average	Fine Grid Average	Instantaneous
New York Department of Environmental Conservation (NYDEC)	UAM-V		NYDEF	
North Carolina Supercomputing Center	MAQSIP			MCNC
Environ/Coordinating Research Council	CAMx	CAMXC	CAMXF	
ICF Consulting/Southern Co.	UAM-V	SAIC	SAIF	

2.4 QUALITY CONTROL

2.4.1 Data Validation

A level of validation is designated by a numeric code indicating the degree of confidence in the data. These levels provide some commonality among data collected and quality controlled by different agencies, and help ensure that all data have received a comparable level of validation. Various data validation levels that apply to air quality and meteorological data have been defined by Mueller and Watson (1982) and Watson et al. (1989). Four levels of data validation are summarized in **Table 9**. Documentation for the observational data used in this study indicates that the data underwent at least Level 1 validation.

Table 9. Data validation levels.

Level	Description
0	Level 0 data validation is essentially raw data obtained directly from the data acquisition systems in the field. Level 0 data have been reduced and possibly reformatted but are unedited and unreviewed. These data have not received any adjustments for known biases or problems that may have been identified during preventive maintenance checks or audits. Routine checks are made during the initial data processing and generation of data, including proper data file identification, review of unusual events, review of field data sheets and result reports, instrument performance checks, and deterministic relationships.
1	Level 1 data validation involves quantitative and qualitative reviews for accuracy, completeness, and internal consistency. Quantitative checks are performed by software screening programs, and qualitative checks are performed by meteorologists or trained personnel who manually review the data for outliers and problems. QC flags, consisting of numbers or letters, are assigned to each datum to indicate its quality. Data are only considered at Level 1 after final audit reports have been issued and any adjustments, changes, or modifications to the data have been made.
2	Level 2 data validation involves comparisons with other independent data sets. This includes, for example, inter-comparing collocated measurements or making comparisons with other measurement systems or analyses. This level is often part of the data interpretation or analysis process.
3	Level 3 validation involves a more detailed analysis when inconsistencies in analysis and modeling results are found to be caused by measurement errors.

2.4.2 Quality Control Codes

The standard quality control codes shown in **Table 10** were used to indicate data quality in the NMI data sets. Observational data were previously subjected to Level 1 data validation and the quality control codes assigned were carried forward into the NMI observational data sets. However, a review of the data indicated that prior data reviewers chose to set values to missing rather than keeping the original data values and flagging them as invalid. In such cases, the flags were not changed because they will not affect the calculation of model performance statistics.

Because model output is not evaluated for quality in the same way as observational data, the special code value of 2 is used to indicate model estimates at site locations.

Table 10. Quality control codes.

QC Code	Definition
0	Valid data
1	Estimated data
2	Model Result
3-6	User defined – not used
7	Suspect data
8	Invalid data
9	Missing data

2.5 FILE NAMES

The names of files included in the NMI database are listed in **Table 11**.

Table 11. NMI database files.

Filename	Contents
NMI_Data_Documentation.doc	This documentation as a Microsoft Word file.
sitelist.txt	List of observation site identifiers and location information.
AQ_observed.dat	Observed ozone, NO_x, and NO_y data (hourly).
HCHO_observed.dat	Observed formaldehyde data (3-hr and 6-hr averages).
CAMx_36km.dat	CAMx coarse grid results at observation sites.
CAMx_12km.dat	CAMx fine grid results at observation sites.
SAI_UAMV_36km.dat	SAI UAM-V coarse grid results at observation sites.
SAI_UAMV_12km.dat	SAI UAM-V fine grid results at observation sites.
MCNC_36km.dat	MCNC MAQSIP coarse grid results at observation sites.
NYDEC_UAMV_12km.dat	NYDEC UAM-V fine grid results at observation sites.

3. REFERENCES

Mueller P.K. and Watson J.G. (1982*), Eastern regional air quality measurements*. Vol. 1, Section 7. Electric Power Research Institute, Palo Alto, CA, Report No. EA-1914.

Watson J.G., Lioy P.J., and Mueller P.K. (1989), *The measurement process: precision, accuracy, and validity*. Air Sampling Instruments for Evaluation of Atmospheric Contaminants, 7th Edition, Hering S.V., ed., American Conference of Governmental Industrial Hygienists, Cincinnati, OH, pp. 51-57.