Water Air Soil Pollut (2011) 222:391–402
DOI 10.1007/s11270-011-0832-7

Quality Assurance Decisions with Air Models: A Case Study of Imputation of Missing Input Data Using EPA's Multi-layer Model

George E. Bowker · Donna B. Schwede · Gary G. Lear · William J. Warren-Hicks · Peter L. Finkelstein

Received: 23 September 2010 / Accepted: 1 April 2011 / Published online: 21 May 2011

Abstract Environmental models are frequently used within regulatory and policy frameworks to estimate environmental metrics that are difficult or impossible to physically measure. As important decision tools, the uncertainty associated with the model outputs should impact their use in informing regulatory decisions and scientific inferences. In this paper, we present a case study illustrating a process for dealing with a key issue in the use and application of air quality models, the additional error in annual mean aggregations resulting from imputation of missing data from model data sets. The case study is based on the US Environmental Protection Agency's Multi-layer Model, which estimates the hourly dry deposition velocity of air pollutants based on hourly measurements of meteorology and site characteristics. A simulation was implemented to evaluate the effect of substituting historical hour-specific average values for missing model deposition velocity predictions on annual mean aggregations. Sensitivity studies were performed to test the effects of different missing data patterns and evaluate the relative impact of the substitution procedure on annual mean SO_2 deposition velocity estimates. The substitution procedure was shown to result generally in long-term unbiased estimates of the annual mean and contributed less than 20% additional error to the estimate even when all data were missing. Consequently, it may be possible to use the historical record of deposition velocities to provide reasonably accurate and unbiased annual estimates of deposition velocities for years without meteorological measurements.

Keywords CASTNET · Deposition · Quality assurance · Decision-making · Model accuracy and precision

G. E. Bowker (✉) · G. G. Lear
Clean Air Markets Division, Office of Air and Radiation,
US Environmental Protection Agency,
1200 Pennsylvania Ave., NW (6204J),
Washington, DC 20460, USA
e-mail: bowker.george@epa.gov

D. B. Schwede
National Exposure Research Laboratory,
Atmospheric Modeling and Analysis Division,
US Environmental Protection Agency,
Mail Drop E243-02, Research Triangle Park,
NC 27711, USA

W. J. Warren-Hicks
EcoStat, Inc.,
P.O. Box 425, Mebane, NC 27302, USA

P. L. Finkelstein
Raleigh, NC 27613, USA

1 Introduction

Quantitative models are frequently used to estimate important environmental parameters. These

estimates provide important information for scientists and policy-makers. However, uncertainty in the model inputs and resulting uncertainty in the model output can reduce confidence in the estimates and diminish the utility of the results. Therefore, guidance on methods and approaches for dealing with key quality assurance issues associated with models, like the effect of missing data on regulatory decision-making and environmental assessment, is required to provide the scientist or decision-maker with a high degree of confidence in using the model outputs. The US Environmental Protection Agency's (USEPA) Office of Air and Radiation, as well as other USEPA programs, frequently use model outputs in support of key regulatory programs. Establishing the quality assurance criteria for using model predictions requires that the USEPA have quantitative metrics describing model uncertainty, that the amount of acceptable uncertainty in the model outputs be stated within the context of the uses and decisions made with the data, and that the resulting quality assurance criteria be thoroughly tested and examined to ensure that the final model predictions are acceptable within the program goals and objectives.

In this paper, we present a case study illustrating a process for imputing missing data in the model data sets. The case study is based on EPA's Multi-layer Model (MLM; Meyers et al. 1998), which is used to estimate dry deposition velocities of acidic compounds and ozone at Clean Air Status and Trends Network (CASTNET) sites (Clarke et al. 1997).

MLM generates hourly predictions of deposition velocity using an Ohm's Law analogy model based on parameterizations of various resistances to deposition, including aerodynamic, boundary layer, stomatal, and cuticular resistances. MLM divides the plant canopy into 21 layers so that within-canopy differences in wind speed, radiation, and leaf area index can be accounted for in the calculation of the deposition velocity. Within MLM, hourly deposition velocity (v_d) estimates are generated using the following:

$$v_d = \left(\frac{1}{\int_0^{h_c} r_c(z)dz + \frac{1}{r_{a,soil}+r_{soil}}} + R_a \right)^{-1} \quad (1)$$

and,

$$r_c(z) = A(z)\left(\frac{1}{r_s(z) + r_b(z) + r_{mes}} + \frac{2}{r_b(z) + r_{cut}} \right) \quad (2)$$

where,

h_c	Height of the canopy (meter)
$r_c(z)$	Canopy resistance (seconds per meter) at height z
$r_{a,soil}$	Subcanopy aerodynamic resistance (seconds per meter)
r_{soil}	Soil resistance (seconds per meter)
R_a	Aerodynamic resistance (seconds per meter)
$A(z)$	Leaf area density at height z
r_s	Stomatal resistance (seconds per meter)
$r_b(z)$	Boundary layer resistance (seconds per meter) at height z
r_{mes}	Mesophyll resistance (seconds per meter), and
r_{cut}	Cuticular resistance (seconds per meter)

The aerodynamic resistance is parameterized as a function of the standard deviation of the wind direction (σ_θ) and the wind speed at 10 m, while the wind speed at the lowest level in the canopy is used to determine the subcanopy aerodynamic resistance. The soil resistance is set to a chemical specific value which is dependent on soil moisture. The canopy resistance is calculated at each level in the canopy. Plant species-specific vertical profiles of leaf area density are used. The stomatal resistance is calculated using the approach of Jarvis (1976), where r_s is determined from a plant species specific minimum stomatal resistance and from factors that account for temperature, soil moisture, and vapor pressure deficit stresses. The boundary layer resistance is calculated from the within-canopy wind profile and the molecular diffusivity of the gas. For the gases currently modeled for CASTNET, the mesophyll resistance is ignored. The cuticular resistance is a chemical specific value that varies with surface wetness. Additional information on the MLM model and the underlying scientific basis for the model structure can be found in Meyers et al. (1998). The ability of the MLM to accurately and precisely predict measured deposition velocities under field conditions is discussed in Finkelstein et al. (2000) and Finkelstein (2001).

For CASTNET, MLM is driven by hourly meteorological measurements taken at on-site 10-m towers. Key site-specific meteorological inputs to the MLM model include hourly measurements of precipitation, wind speed, temperature, standard deviation of wind direction (σ_θ), relative humidity, and solar radiation. Missing hourly metrological data (caused by instrument malfunction, quality assurance checks, laboratory error, etc.) results in a missing MLM model prediction of hourly deposition velocity and therefore a missing value for the hourly deposition flux. The amount of missing data for any MLM model input parameter at any given CASTNET location for a specific year can range from less than 1% to over 90%. Weekly average concentration measurements are made at the CASTNET sites of SO_2, HNO_3, particulate SO_4, NO_3, NH_4, and a suite of base cations, while hourly average concentration measurements are made of O_3. Although not as likely as missing meteorological data, missing concentration values would also result in a missing value for the hourly flux. Once valid hourly fluxes are calculated, the hourly fluxes are aggregated to weekly, quarterly, and then annual deposition estimates. Currently, in the CASTNET program, a data completeness criterion of 69% is used at each step of the process. For example, at least 69% of the hourly deposition fluxes must be present to calculate a valid weekly deposition flux. Therefore, completeness of the flux data at the hourly time step is critical to ultimately providing annual values of deposition. If there are enough missing deposition velocity estimates, the completeness criterion is not met, which reduces the quantity of information available from this valuable monitoring program. Missing data in model input data sets are a major quality assurance issue for all predictive air quality models. Therefore, establishing an approach for resolving missing data issues, designing a procedure for possibly imputing the missing hourly data, and subsequently establishing a criterion for the acceptable amount of missing data are important components of the quality assurance program associated with the model outputs.

Other approaches for imputing missing deposition velocities have been investigated (Lavery et al. 2008) and include substituting particular missing meteorological input parameters using either historical data or data from the nearest CASTNET station. However, replacing missing MLM inputs requires that the model be rerun, adding substantial programmatic computational and procedural expenses. Another alternative approach (Lavery et al. 2008) is to use long-term average

Table 1 Representative CASTNET sites selected for simulation

Site ID	Location	Latitude	Longitude	Plant species
ABT147	Abington, CT	41.84	−72.01	Sugar maple (30%), beech (27%), white oak (20%), grass (12%), Virginia pine (22%), water (5%)
ACA416	Acadia National Park, ME	44.38	−68.26	Mixed wetlands (20%), mixed conifer deciduous (17%), white pine (12%), aspen/birch (11%), beech/maple (9%), spruce (7%), mixed pine (7%), mixed conifer (6%)
CAD150	Caddo Valley, AR	34.18	−93.10	Southern red oak (32%), loblolly pine (24%), grass (24%), water (20%)
DEN417	Denali National Park, AK	63.75	−148.96	Stunted pine (61%), spruce (17%), scrub (9%), mixed deciduous (6%), grass (5%), mixed forest (2%)
DEV412	Death Valley, CA	36.51	−116.85	Rock (95%), sagebrush (5%)
GAS153	Georgia Station, GA	33.18	−84.41	Grass (35%), loblolly pine (27%), apple/peach/pear (13%), maize (10%), wheat (10%), water (5%)
GTH161	Gothic, CO	38.96	−106.99	Grass (45%), aspen (40%), spruce (15%)
NCS415	North Cascades National Park, WA	48.54	−121.45	Rock (36%), white oak (15%), maple (15%), spruce (13%), Virginia pine (10%), grass (8%), pond pine (3%)
PNF126	Cranberry, NC	36.11	−82.05	Grass (38%), chestnut/red oak (32%), maple (30%)
PRK134	Perkinstown, WI	45.21	−90.60	White oak (39%), sugar maple (33%), blue grass (10%), maize (9%), water (9%)
WSP144	Washington Crossing, NJ	40.31	−74.87	White ash (40%), grass (37%), maize (23%)

deposition velocities, but this approach does not capture the diurnal and seasonal patterns in the deposition velocity (Sickles and Shadwick, 2007).

The goals of this study are to develop a method to impute missing MLM deposition velocity estimates and then to quantify the additional uncertainty in the annual average deposition velocity estimate when the method is applied. The method developed was to replace missing deposition velocity estimates with the average value from historical measurements for that hour and day of the year.

2 Methodology

To assess the impact of substituting historical deposition velocity data for missing values in the CASTNET database, an approach was developed to capture the diurnal and seasonal patterns in missing data that could occur across the range of CASTNET sites. Eleven CASTNET sites representative of the geographical and ecological diversity of sites in the CASTNET program were chosen for evaluation (Table 1). The sites ranged from dry ecosystems like site DEV412, a desert location in Death Valley, CA, to sites with high precipitation and large vegetative diversity, like site NCS415 in North Cascades National Park, WA. Many of these sites have been sampled for over 20 years, resulting in an extensive historical record of meteorology and deposition. Of the entire period of record, the years 2000–2005 were selected for further analysis because they had relatively high data completion rates at the selected sites (USEPA 2009). These 66 site–year combinations had 89% valid data, on average, for the 5-year period. Hereafter, the data for these 6 years at the 11 sites are termed the "Master" files. The MLM model generates hourly predictions for SO_2, nitric acid, ozone, and particulate deposition velocity. For this

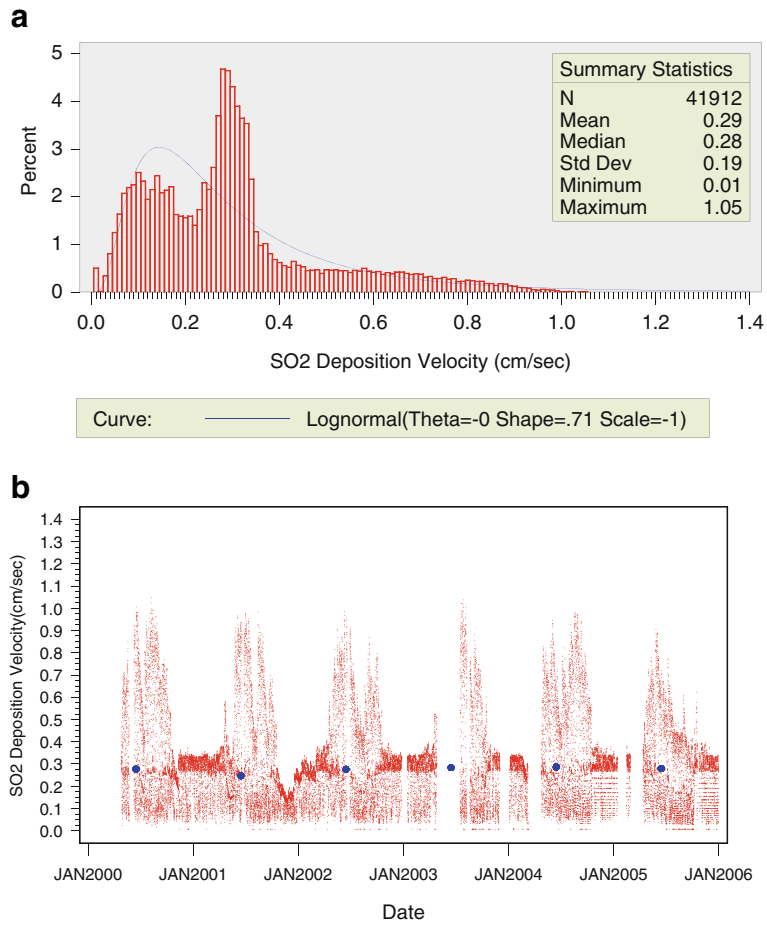

Fig. 1 **a** Distribution of hourly SO_2 deposition velocities at site ABT147 (Abington, CT). **b** Time-series of hourly SO_2 deposition velocities. The *blue dots* indicate the yearly mean SO_2 deposition velocity (averaged over non-missing MLM predictions during the year)

study, SO_2 deposition velocity was chosen for analysis. A comparison of study results among the other deposition outputs showed the greatest variability in annual average deposition velocity was expressed for the SO_2 deposition velocity outputs. Consequently, the study findings presented here represent a "worst-case" scenario, where the general conclusions should apply to the other chemical species measured by CASTNET.

For each of the 11 sites and 6 years, the historical average SO_2 deposition velocity was calculated for each hour of the year using non-missing values from the site's entire historical record with the following exceptions. Values from the first 2 years of operation of the site were excluded from these historical averages to ensure that the soil moisture estimates used in the deposition velocity model had reached equilibrium. The selected year was also excluded from the averages to avoid bias in the evaluation of results. This procedure generated six files for each site (one for each year) which, for this paper, are termed "Historical" files. As an example, the site DEV412 began operation in 1995; the Historical file for the year 2000 for this site would contain one value for each hour of the year (8,760 h), each of which is the average deposition velocity of that hour for the years 1997–1999 and 2001–2005.

The amount and pattern of missing data across CASTNET sites and between years is random and cannot be modeled. Therefore, missing data patterns were simulated by using "Mask" files which contain patterns of missing data during the year 2000 at 78 different CASTNET sites. Tests of the sensitivity of the substitution method to the missing data patterns used were conducted by imposing each of the 78 Mask files on each of the 66 Master files. The resulting missing hours were then replaced by the corresponding hourly average deposition velocity from the Historical file for that site and year. Hourly values missing in the original

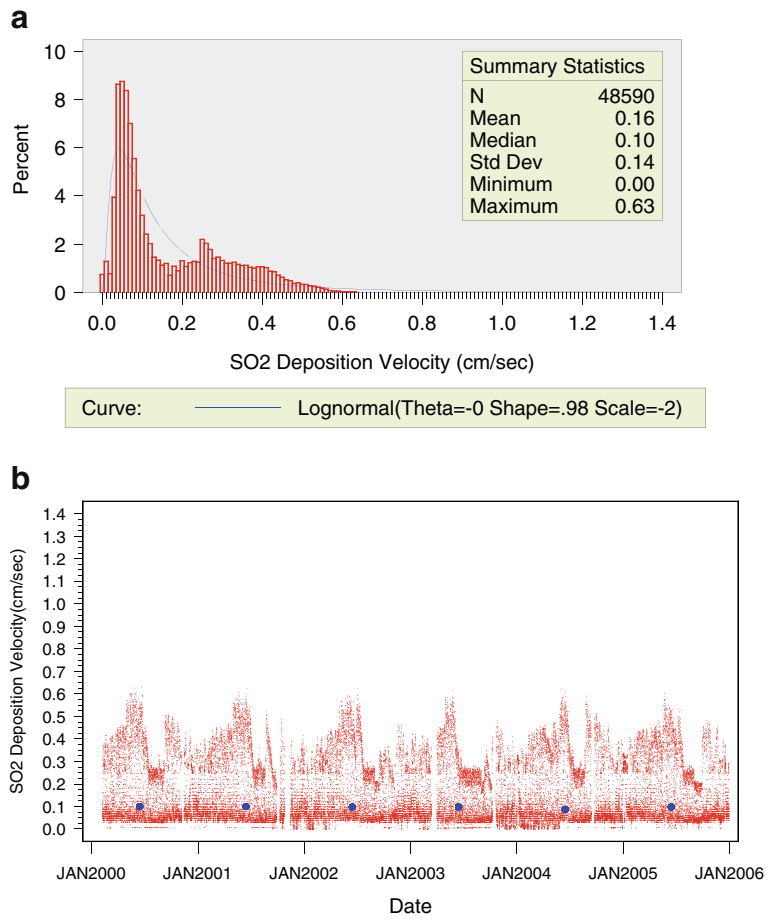

Fig. 2 **a** Distribution of hourly SO_2 deposition velocity at site NCS415 (North Cascades National Park, WA). **b** Time-series of hourly SO_2 deposition velocities. The *blue dots* indicate the yearly mean SO_2 deposition velocity (averaged over non-missing MLM predictions during the year)

Master file were not replaced regardless of whether the value was missing in the Mask file.

Regulatory analyses and environmental assessment decisions typically use aggregated data from CASTNET because of the large geographic extents and long-term temporal scales related to reductions in air emissions mandated by the 1990 Clean Air Act Amendments. Therefore, to be consistent with and relevant to these analyses and assessments, the annual deposition velocity was selected as the metric for evaluating this method.

The annual average SO_2 deposition velocity was calculated for each of the 66 Master files (termed "Master Annual Average"). For each of the 66 Master files, all 78 Mask files were applied, creating patterns of missing data for that file. The new "missing" hourly deposition velocity values were replaced using the Historical file for that site and year. This created 78 "Substitute" files for each of the 66 Master files. We calculated the annual average SO_2 deposition velocity for each of the 78 Substitute files for each of the Master files from the simulation procedure (termed "Substitute annual average"). For each of the 78 data sets (i=1 to 78), a percent difference between the Master Annual Average value and that resulting from the resulting Substitute data files was calculated as:

$$\text{Percent Difference}_i = \left(\frac{(\text{Master Annual Average} - \text{Substitute Annual Average}_i)}{\text{Master Annual Average}}\right) * 100 \quad (3)$$

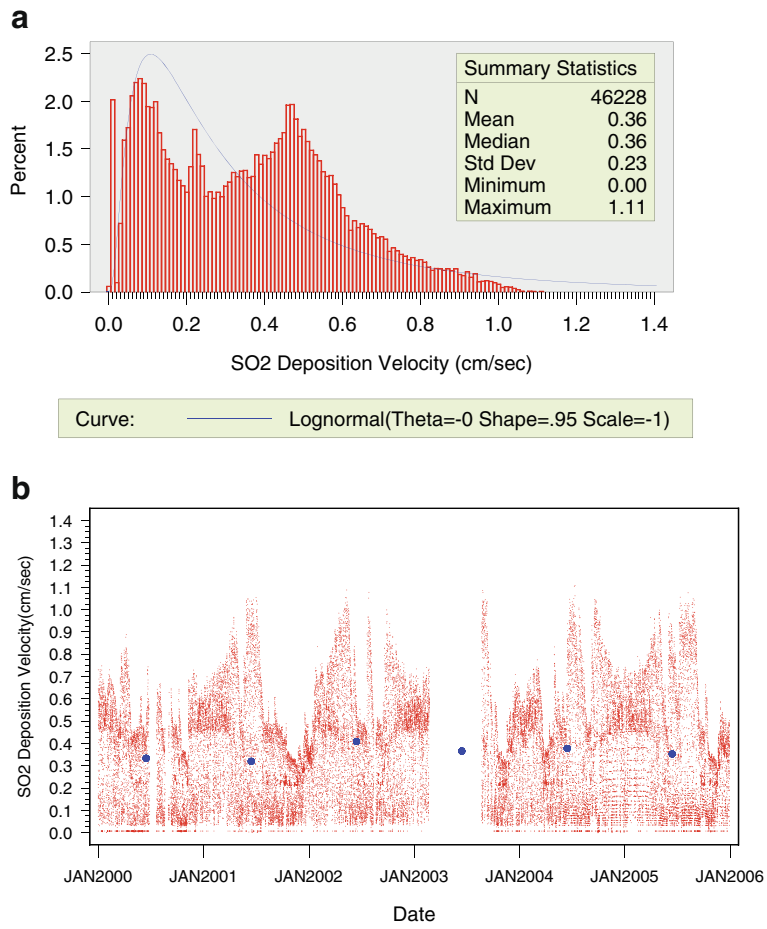

Fig. 3 **a** Distribution of hourly SO2 deposition velocity at site GAS153 (Georgia Station, GA). **b** Time-series of hourly SO2 deposition velocities. The *blue dots* indicate the yearly mean SO2 deposition velocity (averaged over non-missing MLM predictions during the year)

In the above formula, negative values resulted when the annual values after substitution were larger than the annual value obtained from the original data file generated by MLM. Transformation of the Percent Difference_i to Percent Error_i was implemented by taking the absolute difference of each calculated percent difference. A total of 5,148 comparisons were made (11 original Master File sites for the 6 years (2000–2005) times 78 Mask file missing data scenarios). The resulting data were then statistically and graphically evaluated.

3 Results and Discussion

An initial evaluation of the Master file data sets indicates a large variability in the deposition velocity patterns among the sites. Examples of this spatial and ecological variance are shown through comparisons between Figs. 1, 2, 3, and 4, each of which shows hourly deposition velocities for an individual site. For example, in Fig. 1, site ABT147, located in the northeastern United States, clearly shows a seasonal pattern over the 6-year time span used in the simulation. The upper panel (Fig. 1a) presents a histogram of the MLM hourly SO_2 deposition velocity predictions, with a log-normal distribution overlain on the data. Descriptive statistics are presented in the insert. Figure 1b indicates the time-series patterns and trends of the modeled hourly deposition velocity. The blue dots indicate the mean annual SO_2 deposition velocity (averaged over non-missing MLM predictions during the year). The time-series graphic indicates the presence of missing data. Notice that the missing data occur at inconsistent times during any year and are not directly tied to the pattern or degree of temporal variability. Similar information is presented in Figs. 2, 3, and 4 for sites NCS415 (high

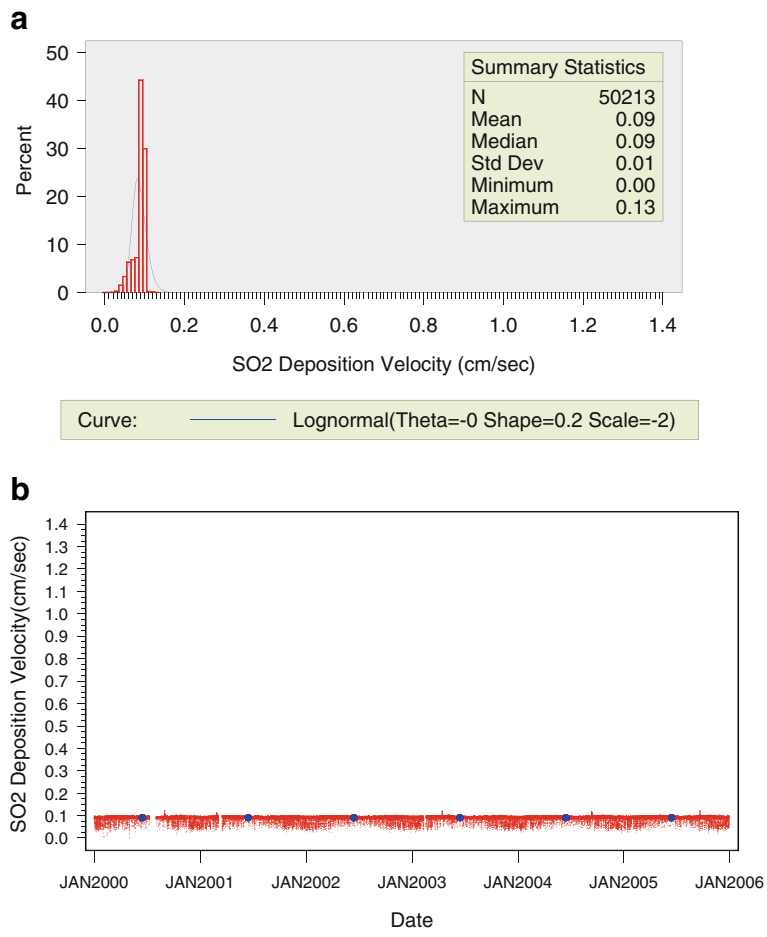

Fig. 4 **a** Distribution of SO_2 deposition velocity at site DEV412 (Death Valley, CA). **b** Time-series of hourly SO_2 deposition velocities. The *blue dots* indicate the yearly mean SO_2 deposition velocity (averaged over non-missing MLM predictions during the year)

elevation site), GAS153 (coastal plain site in the southern United States), and DEV412 (desert site in the western United States). Missing data patterns vary among these example sites, and the patterns, magnitudes, and temporal trends of the MLM SO_2 deposition velocity outputs are highly variable among sites. Generally, sites at lower elevations in the eastern United States have a higher frequency of larger deposition velocity. Vegetation at these sites has a high leaf area index and long active growing seasons. In colder climates (northern and high elevation sites), the plant stomata are open less often resulting in less uptake of gases. In dry regions of the United States, the amount of gaseous chemicals absorbed by vegetation is minimized as plant stomata are often closed to reduce evapotranspiration. Key climatology patterns and ecological factors that affect site-specific deposition velocity include rainfall patterns during the spring and summer seasons, the percentage of site-specific crop and forest coverage, and the site-specific wind speed and turbulence.

The within-year variability of SO_2 hourly deposition velocities in any of the Master files is dependent upon a large number of environmental factors, including vegetation, temporal changes in precipitation, weather patterns, and temperature. Given the range and temporal patterns of deposition velocities illustrated in Figs. 1, 2, 3, and 4, it is clear that accurately estimating a missing hourly MLM prediction from historical data will necessitate the use of hour-specific values rather than a simple annual average value.

A comparison of the range and distribution of deposition velocities for different sites is shown in Fig. 5, which presents the empirical cumulative distribution function (CDF) of non-missing MLM SO_2 hourly deposition velocity predictions for Master files for the years 2000–2005. As was seen through

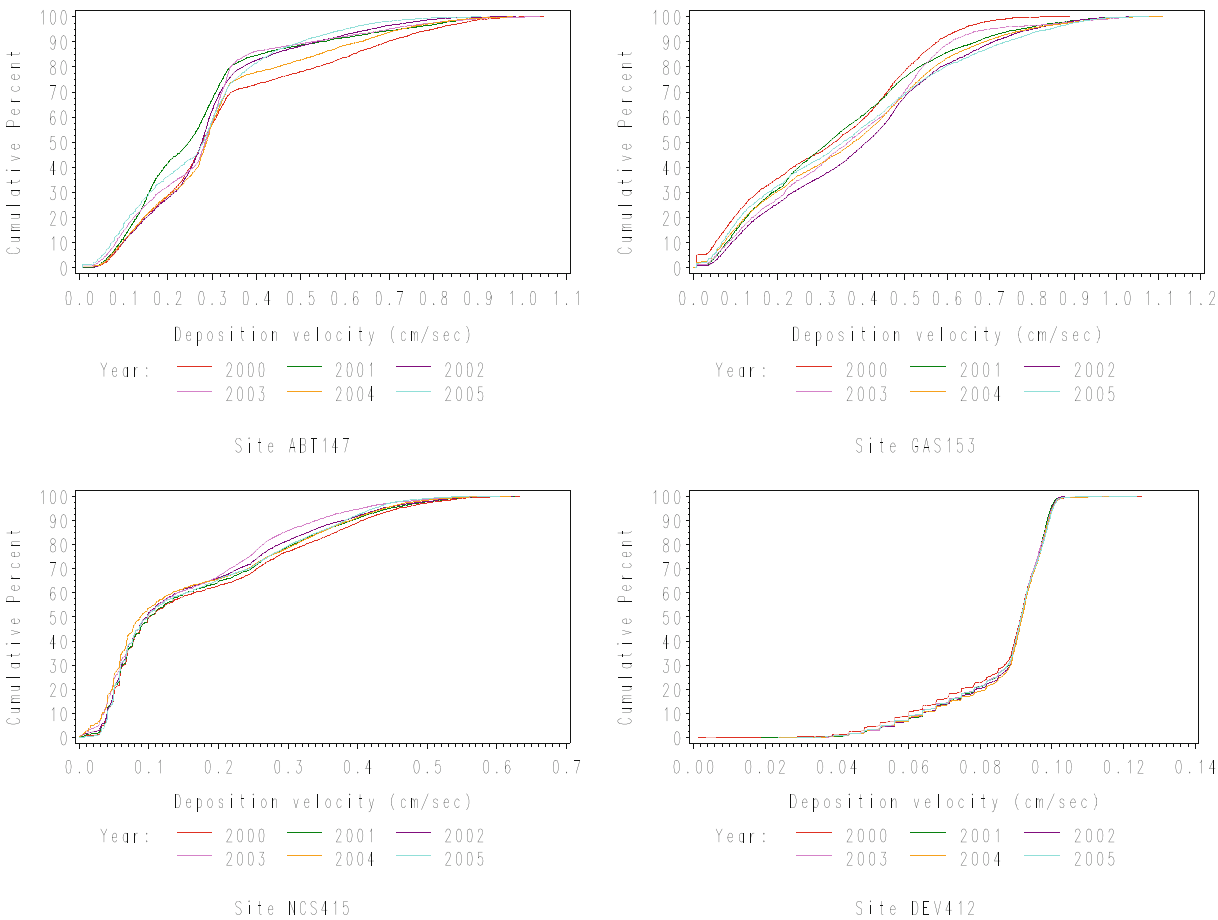

Fig. 5 Cumulative distribution function of hourly SO_2 deposition velocity at selected representative sites

the comparisons of Figs. 1, 2, 3, and 4 and through comparisons of the different panels in Fig. 5, the overall CDF patterns are different between sites. In contrast, the CDF's for different years are relatively consistent for a given site. Examination of the 42 Master files for the seven remaining Master file sites (not shown here) showed results similar to Fig. 5. Therefore, any approach to substitution should incorporate the historical record at the specific site to the degree possible, with longer historical records possibly providing a better estimate of the expected value for any hourly deposition velocity output from MLM.

When the substitution method is applied, the annual average deposition velocity in the Substitute file depends on the year, the number of replaced hours, and on the distribution of the missing values. Sensitivity tests of these factors were conducted by applying the missing data patterns for the 78 Mask files to each of the 66 Master files (11 Master file sites). The valuation of the percent error statistic (n= 5,148 comparisons) resulting from the simulation is displayed in Table 2. The number of missing hours in a year was categorized by bins representing 10% intervals, from 0% to 100% missing hours, and the average percent error for each missing hour bin was computed. The magnitude and distribution of the percent error statistic is site dependent; however, the trend in magnitude of the statistic is generally consistent among the chosen sites. As the percent of missing data increases, the relative error in the average annual value associated with substituting historical hourly average values also increases, with the average relative error increasing from <1% for 1–10% missing hours to 4.0% for 41% to 50% missing hours. It is significant to note that even when 91–100% missing hours were replaced using this method the average relative error in the estimate was only 7.7% and the maximum average relative error was 17.8% (site PRK134).

Figure 6 presents cumulative distributions of the site- and year-specific percent error statistic for selected Master sites. Each line on the graph is generated from the 78 simulations for a single site–year combination. The distributions provide a visual examination of the relative change in percent error between years for a single site. In the examples provided in Fig. 6, and for the other sites not shown, on average the relative change in the error statistic between years was less than 1%. For some sites and

Table 2 Percent error for missing data categories

	Percent missing data categories									
SITE_ID	0–10%	11%–20%	21%–30%	31%–40%	41%–50%	51%–60%	60%–70%	71%–80%	81%–90%	90%–100%
ABT147	0.50	1.52	1.97	3.32	4.62	5.66	5.31	–	–	7.21
ACA416	0.58	1.58	3.23	4.41	5.75	8.85	7.02	9.85	–	11.84
CAD150	0.46	0.93	1.99	2.12	2.82	3.42	4.83	3.47	–	5.10
DEN417	0.93	2.48	4.47	6.23	6.35	8.73	7.67	11.54	–	14.56
DEV412	0.06	0.12	0.18	0.19	0.22	–	0.31	0.40	–	0.54
GAS153	0.61	1.89	3.03	4.29	4.55	–	6.53	6.21	–	7.37
GTH161	0.38	0.95	1.72	1.65	2.02	1.43	2.81	2.52	–	4.49
NCS415	0.42	1.06	1.12	2.19	2.25	1.86	3.58	2.82	–	3.71
PNF126	0.38	1.28	1.76	1.87	2.25	3.24	3.22	–	–	3.33
PRK134	0.93	3.30	5.49	7.43	8.64	10.92	10.73	–	–	17.80
WSP144	0.56	1.71	3.00	2.34	4.40	–	4.94	6.08	–	8.90

Note: Results are not reported when N<5

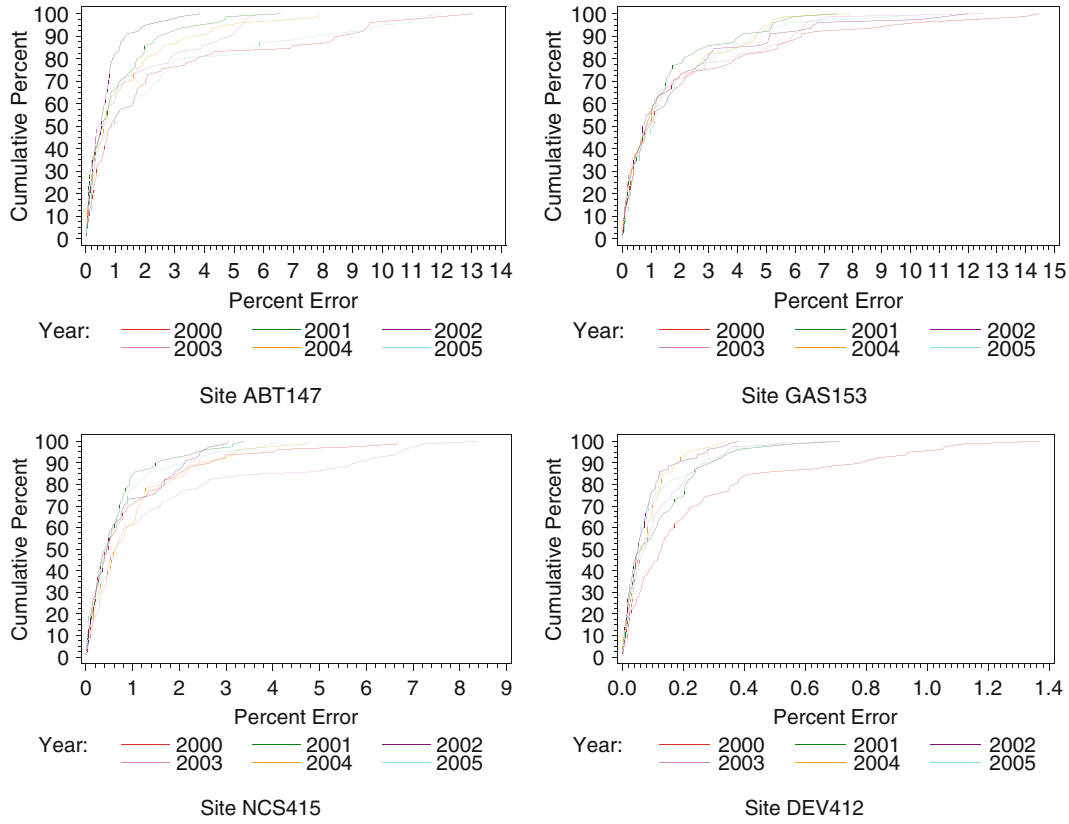

Fig. 6 Cumulative distribution of percent error for representative sites

missing data patterns, the between-year variability is 3–4% percent error, but this occurs infrequently (i.e., in less than 20% of the simulations run). Therefore, the substitution approach results in a relatively consistent expected error over time for a specific site.

This finding leads to two important questions for each specific scientific or programmatic use of the data. First, what magnitude of uncertainty is acceptable for the particular application (trend detection, regulatory decision, or environmental assessment) using the MLM predictions? And second, what is the magnitude of data substitution in the MLM predictions that results in an uncertainty less than that specified in the previous question? A key issue in resolving the above questions is the concept of bias, i.e., whether the substitution process results in estimates that are consistently higher or lower than the known annual mean deposition velocity. The results of the percent difference statistic resulting from the simulation are shown in Table 3. A percent difference of zero indicates no bias, a positive value indicates that on average the Substitute Annual Average is less than the Master Annual Average, and a negative value indicates that on average the Substitute Annual Average is greater than the Master Annual Average. Examination of Table 3 indicates that the tendency for the substitution procedure to result in biased estimates of the annual mean deposition velocity is inconsistent among sites, and trends are not correlated with the amount of missing data. For example, site PRK134 and site ACA416 have negative percent differences along the scale of percent missing data, while sites ABT147 and GAS153 generally are associated with positive percent differences. Other sites have both positive and negative average percent differences across the range of percent missing data. From a CASTNET program perspective, the results are reassuring. For a particular year at any given site, the annual mean deposition velocity resulting from substitution may be either higher or lower than the true mean. That is, there is no general tendency of the substitution procedure to have a large impact on the analysis of deposition trends across time. In addition,

Table 3 Percent difference for missing data categories

	Percent missing data categories									
SITE_ID	0–10%	11%–20%	21%–30%	31%–40%	41%–50%	51%–60%	60%–70%	71%–80%	81%–90%	90%–100%
ABT147	0.10	0.50	0.63	0.48	1.89	1.94	1.24	2.89	–	1.95
ACA416	−0.21	−0.88	−1.53	−1.84	−1.86	−6.09	2.66	-6.19	–	−4.76
CAD150	−0.13	−0.22	−0.66	−1.23	−1.89	−0.70	−2.84	−2.52	–	−2.93
DEN417	−0.01	−0.03	0.11	−0.08	−0.47	0.33	1.37	−1.23	–	1.86
DEV412	−0.03	−0.07	−0.09	−0.17	−0.19	–	−0.31	−0.40	–	−0.46
GAS153	0.15	0.83	0.15	1.33	1.98	–	1.53	1.85	–	1.38
GTH161	0.07	−0.04	0.80	0.28	0.08	−0.79	1.16	0.03	–	1.16
NCS415	0.05	0.00	0.30	−0.07	0.02	−1.03	0.08	−0.13	–	0.83
PNF126	−0.01	−0.35	0.04	0.57	−0.28	−1.37	−0.78	–	–	−0.86
PRK134	−0.38	−1.77	−1.54	−3.80	−3.42	−5.37	−6.41	–	–	−8.66
WSP144	0.04	0.17	0.55	−0.40	−0.95	–	−2.20	1.60	–	−1.45

Note: Results are not reported when $N<5$

examination of Table 3 indicates that for the majority of the sites and missing data categories, the average percent difference is relatively small (<±5%).

4 Conclusions

The objective of this project was to establish quantitative uncertainty relationships for average deposition velocity predictions with varying degrees of missing data from the MLM air quality model. Information on the uncertainty will improve the usefulness of data from CASTNET for environmental decision-making and environmental assessment. A simulation designed to evaluate the effect of substituting missing model predictions on an hourly scale was implemented, and the relative impact of the substitution procedure on the site-specific annual mean SO_2 deposition velocity estimates was evaluated. The substitution procedure was shown to result generally in long-term unbiased estimates of the annual mean. Variations in error resulting from the missing data procedure were shown to be site-specific; however, the interannual variability in the expected error for a given site was generally consistent over time. For most sites, the results of the current study suggests that substitution of numerous hour-specific historical values for missing hourly values leads to only small increases in uncertainty in the resulting annual average SO_2 deposition velocity. Even when all data was missing, the average additional error using this approach was less than 8% and the maximum additional error was less than 20% for the studied sites. Therefore, the use of historical average values which capture the diurnal and seasonal patterns to fill in or substitute for missing model hourly predictions is a reasonable approach for increasing the information content of the publically available CASTNET data.

Results from the current study are helping to guide the EPA's Clean Air Markets Division to develop quantitative relationships between the amount of missing data and uncertainty. These results are being used to help EPA develop quality assurance criteria for the acceptable amount of missing data that can be imputed and still maintain acceptable data quality for its assessments of total deposition.

This article describes the process of developing quantitative metrics of uncertainty for data for use in assessments. While we have illustrated this for data replacement of missing data values, the same con-

cepts apply to instrument or chemical analysis-derived uncertainty and error steps in the collection of the data. Understanding and using quantitative uncertainty/error metrics maximizes the amount of data that is available for use in the particular assessment. For the specific application of CASTNET annual deposition fluxes, this analysis will greatly increase the number of sites (and years) of data that will be acceptable for inclusion in future assessments.

References

Clarke, J. F., Edgerton, E. S., & Martin, B. E. (1997). Dry deposition calculations for the clean air status and trends network. *Atmospheric Environment, 31*(21), 3667–3678.

Finkelstein, P. L. (2001). Deposition velocities of SO_2 and O_3 over agricultural and forest ecosystems. *Water, Air, & Soil Pollution: Focus, 1*(5–6), 49–57.

Finkelstein, P. L., Ellestad, T. G., Clarke, J. F., Meyers, T. P., Schwede, D. B., Hebert, E. O., et al. (2000). Ozone and sulfur dioxide dry deposition to forests: observations and model evaluation. *Journal of Geophysical Research, 105* (D12), 15365–15377. Washington, DC.

Jarvis, P. G. (1976). The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 273*(927), 593–610.

Lavery, T., Rogers, C., Mishoe, K., & Baumgardner, R. E. (2008). Sensitivity analysis of the multilayer model used in the clean air status and trends network (CASTNET). Washington, DC: U.S. Environmental Protection Agency. EPA/600/R-08/126.

Meyers, T. P., Finkelstein, P., Clarke, J., Ellestad, T. G., & Sims, P. F. (1998). A multilayer model for inferring dry deposition using standard meteorological measurements. *Journal of Geophysical Research, 103*(D17), 22645–22661.

Sickles, J. E. II, & Shadwick, D. S. (2007). Seasonal and regional air quality and atmospheric deposition in the eastern United States. *Journal of Geophysical Research, 112*.

U.S. Environmental Protection Agency (2009). Clean air status and trends network. http://www.epa.gov/castnet.