STOCHASTIC DESCRIPTION OF SUBGRID POLLUTANT VARIABILITY IN CMAQ

Jerold A. Herwehe*
NOAA/ARL/Atmospheric Turbulence & Diffusion Division, Oak Ridge, Tennessee
e-mail: Jerry.Herwehe@noaa.gov
Voice: (865) 576-0449 FAX: (865) 576-1327

Jason K. S. Ching and Jenise L. Swall
NOAA/ARL/ASMD on assignment to USEPA/NERL, Research Triangle Park, North Carolina

1. INTRODUCTION

Regional scale air quality (AQ) models are typically limited to relatively coarse resolutions when simulating mean pollutant concentrations for each grid cell volume. However, human exposure and risk assessment require more detailed information on the location and magnitude of hazardous air pollutant, or air toxics, concentrations, with particular interest in capturing extreme values, or "hot spots." Computational fluid dynamics and coupled large-eddy simulation with photochemistry techniques allow AQ simulations with much finer grid spacings, but these types of simulations are impractical for long time integrations or operational use. Thus, some procedure is needed to represent the subgrid pollutant concentration extremes in regional models without requiring concurrent fine resolution simulations.

A methodology and software are being developed to perform statistical analyses on available fine resolution gridded model results in order to quantify the subgrid pollutant variability not represented in current regional AQ models. Our specific goal is to utilize Community Multiscale Air Quality (CMAQ) (see http://www.epa.gov/asmdnerl/models3/cmaq.html and Ching and Byun 1999) simulation output to determine pollutant probability density function (pdf) characteristics and parameters for input into the Hazardous Air Pollutant Exposure Model (HAPEM) (see http://www.epa.gov/ttn/fera/human_hapem.html).

2. APPROACH

Fine resolution CMAQ results have been used as sample data during the development of our methodology for quantifying subgrid pollutant variability. To treat the sample data with complete objectivity, the Exploratory Data Analysis (EDA) approach was used (NIST/SEMATECH 2003). EDA emphasizes numerous graphical techniques, along with several quantitative techniques, to reveal the underlying structure of the sample data set. For performing EDA, the National Institute of Standards and Technology (NIST) has made freely available a companion statistical analysis software package called Dataplot (Filliben 1982, 1984) for interactive data exploration, but which also supports a scripting capability for more complex automated tasks (see http://www.itl.nist.gov/div898/software/dataplot/). The Dataplot command script under development for this research has been dubbed Concentration Distribution Function –ware, or CDFware.

For each sample pollutant concentration data set, CDFware conducts numerous statistical tests and produces various graphical and text outputs in order to objectively determine the best-fit univariate distribution pdf which represents the subgrid pollutant concentration variability. CDFware produces a summary table which provides copious statistical quantities. A standard 4-plot analysis is created (a run sequence plot, a lag plot, a histogram, and a normal probability plot). CDFware next produces a bootstrap plot to indicate the best location parameter (mean, median, or midrange), followed by a runs test and an autocorrelation plot to check randomness. A Tukey-Lambda probability plot correlation coefficient (PPCC) plot is generated to indicate the best symmetrical distribution family that might fit the data. CDFware then checks for a uniform distribution using a uniform probability plot fit criterion and for a normal distribution using the Anderson-Darling test. The presence of outliers is checked using Grubbs' test. If the data are nearly uniform, then CDFware produces a relative histogram plot with a fitted uniform pdf, plus a uniform probability plot. If the data are approximately normal, then a relative histogram with fitted normal pdf plot is created along with a normal probability plot. If neither uniform nor normal, CDFware makes one last symmetric distribution check by creating logistic distribution PPCC and probability plots if the Tukey-Lambda shape parameter is less than 0.05.

If the sample data set has been determined to be asymmetrical and the skewness is positive, CDFware currently tests for a best-fit from these ten right-skewed distributions: Weibull, lognormal, gamma, power normal, power lognormal, skewed normal, Frechet, generalized extreme value, inverted Weibull, and chi-squared. If the data skew is negative, CDFware currently tests for a best-fit from two left-skewed extreme value distributions: Weibull and Frechet. Two-iteration PPCC plots, the appropriate probability plot, and a relative histogram plot with a fitted distribution pdf are produced for each distribution. CDFware then automatically compares the maximum PPCC values to determine the best distribution fit (currently selecting the distribution with the highest maximum PPCC value). The CDFware analysis finishes by producing a report summarizing the statistical findings and the chosen distribution pdf with parameters, plus a concise output data file for use as a source in building input files for other plotting packages.

* *Corresponding author address:* Jerold A. Herwehe, NOAA/ATDD, 456 S. Illinois Ave., P.O. Box 2456, Oak Ridge, TN 37831-2456.

3. EXAMPLE RESULTS

Ground-level pollutant concentration mixing ratios from the 1.33 km-spaced nested grid of a CMAQ simulation on 14 July 1995 for the greater Philadelphia, Pennsylvania, area were used during the development of the subgrid concentration variability methodology and the CDFware postprocessing package. Any fine resolution model output could have been statistically analyzed, but our goal is to provide a link between the CMAQ-generated air toxics concentrations and HAPEM.

3.1 *Sample Data from a CMAQ Case Study*

Surface acetaldehyde (ALD2) mixing ratios for 15:00 LST on 14 July 1995 are shown in Fig. 1 at two different grid resolutions. Figure 1a shows the mean ALD2 mixing ratio taken directly from the finest nested CMAQ grid with 1.33 km spacing. There are 99×99 of the (1.33 km)^2 cells in Fig. 1a, with the southwest corner at 39.4667° N and 76.0147° W, and the northeast corner at 40.4424° N and 74.1507° W. Figure 1b shows the mean ALD2 mixing ratio for 10×10 (12 km)^2 grid cells derived by averaging 9×9 blocks of the 1.33 km output. As expected, the details and extreme values seen in Fig. 1a are lost in the averaging shown in Fig. 1b. Though present to varying degrees at all grid resolutions in Eulerian models, clearly the averaging inherent in relatively coarse grid regional air quality models often results in an inadequate representation of extreme concentrations needed for exposure models.

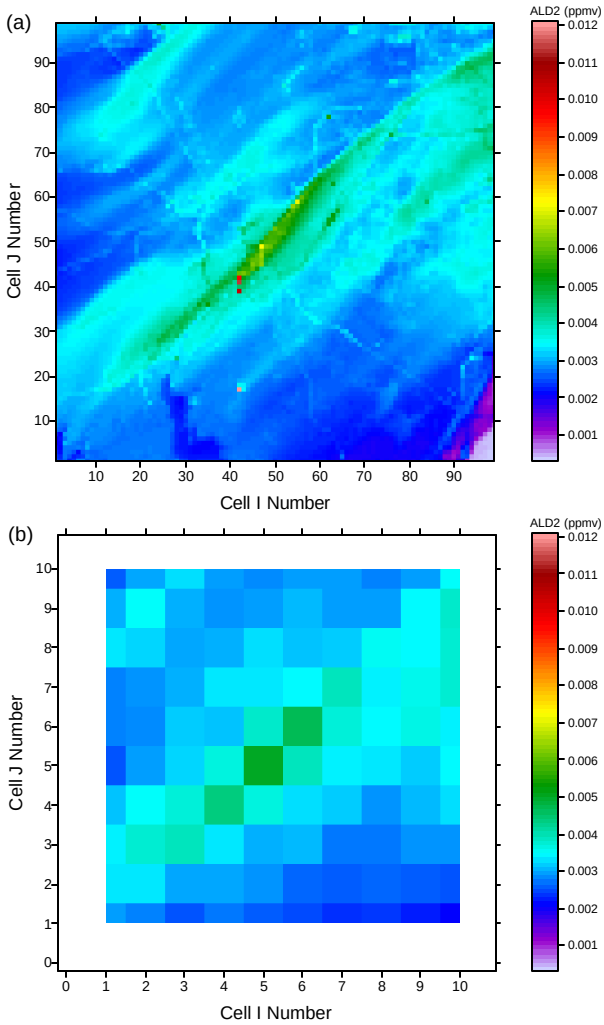

FIG. 1. Mean surface mixing ratio of acetaldehyde (ALD2) for 15:00 LST 14 July 1995 from a CMAQ simulation of the Philadelphia, PA, area shown at (a) 1.33 km grid spacing and (b) 12 km grid spacing. Area coverage and color spectrum range are set to the same scale to facilitate comparison.

3.2 *CDFware Analysis of Acetaldehyde*

Each (12 km)^2 grid cell shown in Fig. 1b, with each consisting of a set of 81 randomized "sample data" from the original 1.33 km grid, was analyzed using the CDFware distribution analysis program according to the procedure described in section 2.

Figure 2 shows the final best-choice distribution types (in no particular order) for the 15:00 LST subgrid ALD2 mixing ratios at 12 km grid spacing as determined by the CDFware subgrid concentration variability analysis package. A wide variety of distributions and no discernible pattern can be seen in these results. The number of cells for each distribution type in this case is: uniform 12, normal 13, Weibull for positive skew 15, lognormal 4, gamma 3, power normal 10, power lognormal 6, skewed normal 13, Frechet for positive skew 6, generalized extreme value 3, inverted Weibull 2, chi-squared 1, Weibull for negative skew 10, and Frechet for negative skew 2.

For skewed distributions, CDFware currently chooses the best-fit distribution based solely on the maximum PPCC value. In practice, there is usually not much difference between the maximum PPCC values of the top few choices for each data set, meaning several distributions may produce nearly equally good fits.

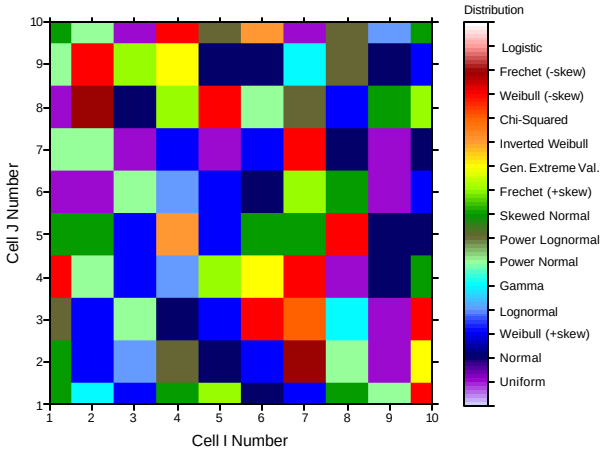

FIG. 2. Map of best-choice distribution type for each (12 km)^2 grid cell as determined by CDFware for acetaldehyde at 15:00 LST 14 July 1995 from the CMAQ simulation.

3.3 *Weibull-Only Analysis of Acetaldehyde*

The chaotic arrangement of distributions shown in Fig. 2 and the small differences between distribution maximum PPCC values for a given cell motivated a subjective approach to the subgrid concentration distribution analysis. Because the Weibull distribution accounted for the largest share (25%) of the CDFware-chosen distributions for the 15:00 LST ALD2, the assumption was made that a Weibull distribution could be successfully applied to the entire domain and possibly yield patterns in pdf parameters that would permit development of parameterizations for subgrid pollutant concentrations.

Figure 3 shows relative histograms for ALD2 for an area around central Philadelphia cell (I05, J05) with the fitted Weibull pdf curves overlaid. Weibull fits for 12 km cells (I05, J03) and (I04, J06) are fairly good, while other Weibull pdf fits are rather poor, such as for cells (I04, J04) and (I05, J05). Even so, Weibull distributions do a reasonable job of representing the extreme values present in some of these sample ALD2 data.

For each fitted distribution, in addition to the maximum PPCC value, CDFware determines all parameters needed to construct the probability density function. Maps of these values for the 15:00 LST acetaldehyde Weibull-only analysis are shown in Fig. 4.

The goodness-of-fit for each Weibull distribution can be gauged by examining Fig. 4a. Most maximum PPCC values are above 0.98, but as expected, a few $(12\ km)^2$ cells stand out as relatively poor Weibull fits,

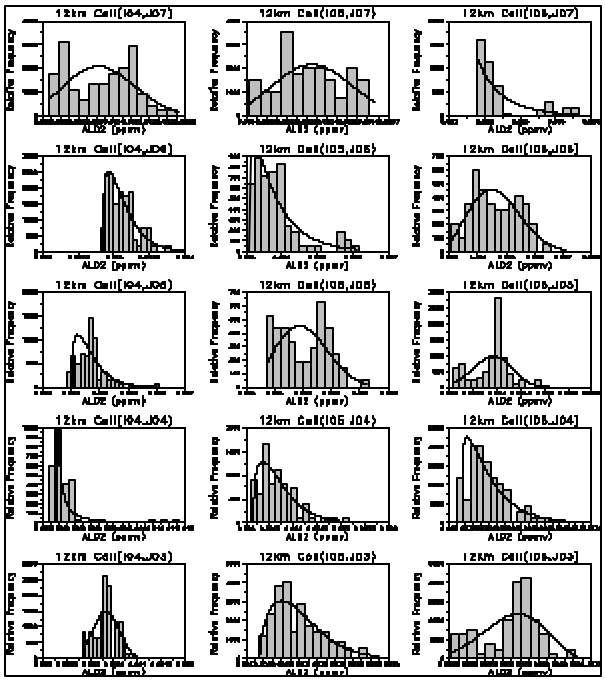

FIG. 3. Relative histograms (relative frequency versus mixing ratio) of surface acetaldehyde with fitted Weibull probability density functions (heavy line) for selected $(12\ km)^2$ cells centered on central Philadelphia cell (I05, J05) at 15:00 LST 14 July 1995 from the CMAQ simulation.

thus implying that a different distribution model would be more appropriate for these cells.

Figures 4b-d illustrate the individual parameters used in the general form of the Weibull distribution pdf (Bury 1999), shown in Eq. (1) for the minimum order statistic (for positive skew):

$$f(x;\, m, s,\, l) = \frac{l}{s}\left(\frac{x-m}{s}\right)^{l-1} \exp\left\{-\left(\frac{x-m}{s}\right)^{l}\right\} \qquad (1)$$

for $x, m \geq 0$ and $s,\ l > 0$

where l is the shape parameter, m is the location parameter, and s is the scale parameter. The location parameter locates the model *f* on its measurement axis, which for a Weibull distribution is not the same as the mean for a normal distribution. This fact can be verified by comparing Figs. 4c and 1b (even though the tile color ranges are different). Likewise, the Weibull scale parameter is not the same as the standard deviation of a normal distribution, though both denote the relative horizontal stretching or contracting of the distribution.

The Weibull shape parameter values shown in Fig. 4b reveal no particular pattern, but the 12 km cells containing the acetaldehyde point sources, cells (4, 2) and (4, 4), distinctly show low shape parameter values appropriate for long-tailed distributions. The Weibull location parameters of Fig. 4c display just a hint of the southwest-northeast structure of the ALD2 mixing ratio field of Fig. 1a. The Weibull scale parameters in Fig. 4d have mostly randomly placed small values, except for the SW-NE "plume" seen starting from the downtown Philadelphia cell (5, 5) which has relatively large values of s. Adapting these results into a new subgrid pollutant variability parameterization would be difficult.

4. CONCLUSIONS

CDFware analyses of the CMAQ model results for acetaldehyde (ALD2) at 15:00 LST on 14 July 1995 from the Philadelphia case study were presented here and the current findings generally did not reveal any discernible pattern or order. Other CMAQ pollutants have been analyzed with CDFware and have also yielded essentially inconclusive results. The CDFware distribution fitting works well and its results can still enhance the input stream to the human risk and exposure models, especially for the higher resolution urban census tract scales. But the desire to utilize CDFware analyses results to develop parameterizations of subgrid pollutant variability for use in coarse grid regional air quality models remains unfulfilled for now.

This research is a work in progress and development continues on improving the Concentration Distribution Function -ware (CDFware) subgrid pollutant concentration variability analysis program. Desired CDFware improvements include the ability to detect multimodal (particularly the relatively common bimodal) data and to fit mixed multiple distributions, such as a mixture of two Weibull distributions in the bimodal case, to the data set. CDFware will also be applied to higher resolution output from neighborhood models in order to determine whether more coherent parameter fields can

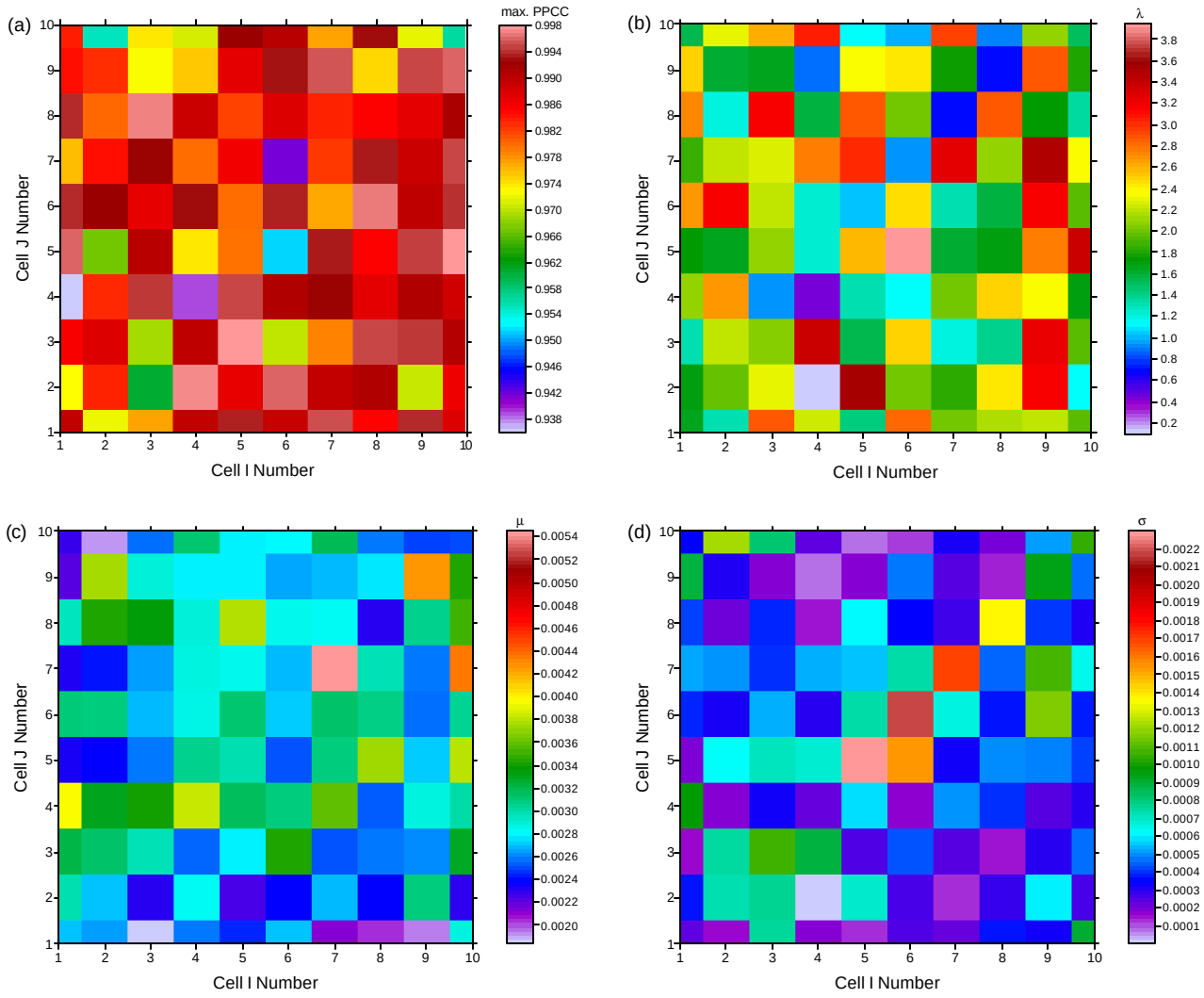

FIG. 4. Weibull probability density function parameters generated by a special Weibull-only version of CDFware for acetaldehyde at 15:00 LST 14 July 1995 from the CMAQ Philadelphia study. Shown are (a) the maximum probability plot correlation coefficient (PPCC) value, (b) the Weibull shape parameter l, (c) the Weibull location parameter m, and (d) the Weibull scale parameter s for each $(12\ \text{km})^2$ grid cell.

be detected at the finer resolutions.

Acknowledgments: This research was supported by the National Oceanic and Atmospheric Administration's Air Resources Laboratory and the U.S. Environmental Protection Agency's National Exposure Research Laboratory. Disclaimer: *This paper has been reviewed in accordance with the United States Environmental Protection Agency's peer and administrative review policies and approved for presentation and publication.*

5. REFERENCES

Bury, K., 1999: *Statistical Distributions in Engineering*. Cambridge University Press, 362 pp.

Ching, J., and D. Byun, 1999: Introduction to the Models-3 framework and the Community Multiscale Air Quality model (CMAQ). In *Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System*, edited by D. W. Byun and J. K. S. Ching, EPA-600/R-99/030, Chapter 1, National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Filliben, J., 1982: Dataplot — An interactive high-level language for graphics, non-linear fitting, data analysis, and mathematics. *Proceedings of the Third Annual Conference of the National Computer Graphics Association*, Anaheim, CA.

——, 1984: Dataplot introduction and overview. NBS Special Publication 667, U.S. Department of Commerce, 112 pp.

NIST/SEMATECH, cited 2003: *NIST/SEMATECH e-Handbook of Statistical Methods*. [Available online at http://www.itl.nist.gov/div898/handbook/.]