QUANTIFYING SUBGRID POLLUTANT VARIABILITY IN EULERIAN AIR QUALITY MODELS

Jerold A. Herwehe*

NOAA/ARL/Atmospheric Turbulence & Diffusion Division, Oak Ridge, Tennessee

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 Eulerian air quality (AQ) models are typically limited to relatively coarse grid resolutions when simulating mean pollutant concentrations for each grid cell volume. However, emergency management, human exposure and risk assessment require more detailed information on the location and magnitude of hazardous air pollutant, or air toxics, concentrations, with a particular interest in capturing extreme values, or "hot spots." Though continuous advancements in computing power and improvements in nested grid techniques have allowed regional scale air quality models to simulate down to one kilometer grid spacing, this resolution is inadequate in urban areas for human exposure modeling based on census tracts. Developments in neighborhood scale modeling using computational fluid dynamics (CFD) and coupled largeeddy simulation (LES) with photochemistry techniques allow AQ simulations with grid spacings of meters to tens of meters for domains limited to several kilometers, but these types of simulations are impractical for long time integrations or operational use. To bridge the gap between regional and neighborhood scale AQ models, some procedure is needed to represent the subgrid extreme pollutant concentrations in regional models without requiring concurrent fine resolution simulations.

The purpose of the present research is to develop a methodology and associated software tools to perform statistical analyses on available fine resolution gridded model results in order to quantify the subgrid pollutant variability not represented in current regional air quality models. Desired products include, but are not limited to, pollutant probability density functions (pdfs) for use in human exposure models and new parameterizations to represent subgrid pollutant variability in regional air quality prediction systems. Our current specific goal is determine pollutant pdf characteristics and to parameters to be used as input to the Hazardous Air Exposure Model Pollutant (HAPEM) (see http://www.epa.gov/ttn/fera/human hapem.html).

2. APPROACH

Relatively fine resolution model results have been used as statistical 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 to the analyst the underlying structure of the sample data set. In essence, the data lead the analysis; no presumptions about the data are made. 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). Though primarily designed for interactive data exploration, Dataplot also supports a powerful scripting capability for more complex automated tasks. For this reason, the fact that Dataplot runs on most computing platforms (e.g. Unix, Linux, Windows), and the availability of the well-tested Dataplot statistical routines as open-source Fortran77 code, Dataplot was chosen as the base tool for developing the subgrid pollutant variability analysis methodology of the present (Dataplot information can be found at research. 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. Currently, CDFware uses the EDA approach to systematically analyze fine resolution gridded model results to objectively determine best-fit univariate distributions which represent subgrid pollutant concentration variability. For each sample pollutant concentration data set, CDFware conducts numerous statistical tests and produces various graphical and text outputs on its way to determining the best distribution pdf to fit the data. CDFware first produces a summary table which provides copious statistical quantities (midrange, mean, median, standard deviation, etc., plus various quantiles, moments, and probability plot correlation coefficients). Then for a quick visual overview of the data set, a standard 4-plot analysis is produced, which consists of a run sequence plot (to check for outliers and for drift in the sample location and its variation), a lag plot (to check data randomness), a histogram (to view the data distribution shape), and a normal probability plot (which is a straight line if the data are normally distributed). CDFware next produces a bootstrap plot to find the best location parameter for the data (mean, median, or midrange), followed by a runs test and an autocorrelation plot to check randomness (i.e., a lack of spatial or temporal correlation within a sample set). A

^{*} Corresponding author address: Jerold A. Herwehe, NOAA/ATDD, 456 S. Illinois Ave., P.O. Box 2456, Oak Ridge, TN 37831-2456; e-mail: Jerry.Herwehe@noaa.gov.

two-iteration Tukey-Lambda probability plot correlation coefficient (PPCC) plot is generated next to indicate the best symmetrical distribution family that might fit the data. To eliminate the simplest distribution cases first, CDFware then checks for a uniform distribution using a uniform probability plot fit criteria and for a normal distribution using the Anderson-Darling test. Data variation drift is checked next using the Bartlett test for the normal case and the Levene test for nonnormal data. Based on the answers to several of these tests, CDFware determines whether the sample is under "statistical control." The presence of outliers is checked next using Grubbs' test. If the data were determined to be nearly uniform, then CDFware produces a relative histogram plot with a fitted uniform pdf, plus a uniform probability plot with a fitted line. If the data were found to be approximately normal, then a relative histogram with fitted normal pdf plot is created along with a normal probability plot with fitted line. 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.

By this point the sample data set has been determined to be asymmetrical. If the skew is positive, CDFware currently tests for a best-fit from these ten right-skewed distributions: Weibull (extreme value distribution based on the minimum order statistic), lognormal, gamma, power normal, power lognormal, skewed normal, Frechet (type II extreme value distribution using the maximum order statistic), generalized extreme value, inverted Weibull, and chisquared. If the data skew is negative, CDFware currently tests for a best-fit from two left-skewed extreme value distributions: Weibull (based on the maximum order statistic) and Frechet (based on the minimum order statistic). 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.

Several underlying assumptions must be met in order for a univariate statistical model to be valid: 1) the sample data are uncorrelated with one another, 2) the random component has a fixed distribution, 3) the deterministic component consists of only a constant, and 4) the random component has fixed variation (NIST/SEMATECH 2003). As described earlier, CDFware checks that these assumptions are satisfied and records its findings in a data set summary report.

3. EXAMPLE RESULTS

Output from the 1.33 km-spaced nested grid of the Community Multiscale Air Quality (CMAQ) modeling system (Ching and Byun 1999) is being used during the

development of the subgrid concentration variability methodology and the CDFware postprocessing package. (Additional CMAQ information can be found at <u>http://www.epa.gov/asmdnerl/models3/cmaq.html.</u>) Any fine resolution model output could have been statistically analyzed, but our focus in a related research project is on creating a link between the CMAQprovided air toxics concentrations and HAPEM (Ching et al. 2004). All CDFware results discussed here are derived from the 1.33 km grid ground-level pollutant concentration output of a CMAQ simulation of the Philadelphia, Pennsylvania, area for 14 July 1995.

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



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.



FIG. 2. Histograms (counts versus mixing ratio) of surface acetaldehyde for each (12 km)² cell at 15:00 LST 14 July 1995 from the CMAQ simulation. Cell (101, J01) is the southwest corner and cell (110, J10) is the northeast corner. Plot *x*-axes used automatic ranging to enhance distribution shape clarity.

ALD2 mixing ratio taken directly from the CMAQ output at the finest nested grid of 1.33 km grid spacing. There are 99×99 of the $(1.33 \text{ km})^2$ cells in Fig. 1a, with the southwest corner located at 39.4667° N latitude and 76.0147° W longitude, 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 from the averaging of 9×9 blocks of the 1.33 km output. Scales for the area coverage and mixing ratio range were kept the same to facilitate comparison of Figs. 1a and 1b. As expected, the details and extreme values (such as point sources and other "hot spots") 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.

The pollutant subgrid concentration variability missing in Fig. 1b is illustrated by the ALD2 histograms of Fig. 2 for each $(12 \text{ km})^2$ grid cell. The histograms for 12 km cells (I04, J02) and (I04, J04) reflect the presence of the bright red to pink ALD2 point sources seen in Fig. 1a. Note the general variety of distribution shapes, with no obvious pattern to the layout. Both spatially and temporally, this sort of distribution complexity exists, more or less, for the other investigated CMAQ pollutants such as carbon monoxide (CO), ozone (O₃), nitric oxide (NO), nitrogen dioxide (NO₂), and formaldehyde (FORM).

3.2 CDFware Analysis of Acetaldehyde

In an attempt to sort out the complexity of Fig. 2, the CDFware distribution analysis program was applied to the 15:00 LST acetaldehyde mixing ratio data. Each $(12 \text{ km})^2$ grid cell, each consisting of a set of 81 randomized "sample data" from the original 1.33 km grid, was analyzed according to the procedure described in section 2.

An example 4-plot analysis from CDFware for 12 km cell (I08, J03) appears in Fig. 3. A quick inspection of the four subplots reveals that the data are free from drift, random, have a positive skew, and are nonnormal. A 4-plot analysis is produced for each grid cell data set and provides a visual summary of the data set characteristics.

The Tukey-Lambda shape parameter λ for the 15:00 LST ALD2 concentrations for each (12 km)² grid cell is shown in Fig. 4. The shape parameter provides guidance as to which family of symmetrical distributions the data may belong according to the following (NIST/SEMATECH 2003): $\lambda = -1$ for an approximately Cauchy distribution, $\lambda = 0$ for exactly logistic, $\lambda = 0.14$ for approximately normal, $\lambda = 0.5$ for a U-shaped distribution, and $\lambda = 1$ for an exactly uniform distribution. Tukey-Lambda shape parameters less than 0.14 indicate increasingly heavy or long-tailed distributions as λ goes to -1, while shape parameters larger than 0.14 indicate shorter-tailed distributions. The two 12 km cells that contain the ALD2 point sources, (4, 2) and (4, 4),



FIG. 3. Example 4-plot analysis from CDFware for acetaldehyde at 15:00 LST 14 July 1995 for $(12 \text{ km})^2$ cell (8, 3) from the CMAQ simulation.

stand out as long-tailed distributions in Fig. 4 due to the interpretation of the sources as outliers. Several cells, including the downtown Philadelphia cell at (5, 5), are shown to be approximately uniform and generally agree with several of the corresponding histograms of Fig. 2. However, the downtown cell, for example, is actually bimodal. This test shows that the Tukey-Lambda shape parameter loses value as a distribution family indicator when the data set is not unimodal nor symmetrical.

An example CDFware-generated PPCC plot set for the positively skewed ALD2 concentration distribution of cell (I08, J03) is shown in Fig. 5. Of the ten distribution types tested against the data, this gamma distribution produced the closest fit as judged by its having the largest maximum PPCC value (0.995588 after two iterations) and supported by the reasonably good fit to



FIG. 4. Tukey-Lambda shape parameter for acetaldehyde for each $(12 \text{ km})^2$ cell at 15:00 LST 14 July 1995.



FIG. 5. Example probability plot correlation coefficient (PPCC) plot set from CDFware for acetaldehyde at 15:00 LST 14 July 1995 for $(12 \text{ km})^2$ cell (8, 3).

the straight line in the gamma probability plot. The gamma probability density function curve is also seen as having a good fit to the relative histogram of the sample data.

Figure 6 shows the final best-choice distribution types (in no particular order) for the 15:00 LST subgrid acetaldehyde mixing ratios at 12 km grid spacing as determined by the CDFware subgrid concentration variability analysis package. Quite a bit of distribution variety and no discernible pattern can be seen in these results. The resulting 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



FIG. 6. Map of best-choice distribution type for each $(12 \text{ km})^2$ grid cell as determined by CDFware for acetaldehyde at 15:00 LST 14 July 1995 from the CMAQ simulation.

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. For example, the gamma distribution shown earlier for 12 km cell (I08, J03) had a maximum PPCC value of 0.995588, but chisquared, lognormal, and Weibull distribution fits for the same cell yielded maximum PPCC values of 0.995585, 0.995367, and 0.995044, respectively, meaning these are all nearly equally good fits to the data.

3.3 Weibull-Only Analysis of Acetaldehyde

The seeming chaotic arrangement of distributions shown in Fig. 6 and the often 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. Perhaps by choosing a single distribution model such as Weibull, patterns in the pdf parameters may emerge that will assist in the development of parameterizations for subgrid pollutant concentrations.

Figure 7 shows the relative histogram version of Fig. 2 for acetaldehyde, this time with the fitted Weibull pdf curves overlaid. Some of the Weibull fits are quite good, such as for 12 km cells (I03, J08), (I05, J03), and (I10, J05). Other Weibull pdf fits are not so good, such as for cells (I01, J04), (I02, J10), and (I06, J07). In general, though, the Weibull distribution does 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. 8.

The goodness-of-fit for each Weibull distribution can be gauged by examining Fig. 8a. Most maximum PPCC values are above 0.98, but as expected, a few $(12 \text{ km})^2$ cells stand out as relatively poor Weibull fits, thus implying that a different distribution model would be more appropriate for these cells.

Figures 8b-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):

$$f(x; \mu, \sigma, \lambda) = \frac{\lambda}{\sigma} \left(\frac{x - \mu}{\sigma} \right)^{\lambda - 1} \exp \left\{ - \left(\frac{x - \mu}{\sigma} \right)^{\lambda} \right\}$$
(1)

• >

for $x, \mu \ge 0$ and $\sigma, \lambda > 0$

where λ is the shape parameter, μ is the location parameter and σ 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. 8c and 1b (even though the tile color ranges are different). Likewise, the Weibull scale parameter is not the same as the standard deviation of



FIG: 7. Relative histograms (relative frequency versus mixing ratio) of surface acetaldehyde with fitted Weibull probability density functions (heavy line) for each (12 km)² cell at 15:00 LST 14 July 1995 from the CMAQ simulation. As in Fig. 2, cell (101, J01) is the southwest corner and cell (110, J10) is the northeast corner.



FIG. 8. 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 λ , (c) the Weibull location parameter μ , and (d) the Weibull scale parameter σ for each (12 km)² grid cell.

a normal distribution, though each denote the relative horizontal stretching or contracting of the distribution.

The Weibull shape parameter values shown in Fig. 8b 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 magnitudes. The Weibull location parameters of Fig. 8c display just a hint of the southwest-northeast structure of the ALD2 mixing ratio field of Fig. 1a. And finally, the Weibull scale parameters in Fig. 8d 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 σ . Adapting these results into a new subgrid pollutant variability parameterization would be difficult.

4. CONCLUSIONS

Results from 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 as pertains to 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 refining and improving the Concentration Distribution Function -ware (CDFware) subgrid pollutant concentration variability analysis program. Desired CDFware improvements would include the ability to detect multimodal (particularly the relatively common bimodal) data distributions 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-scale models in order to determine whether more coherent parameter fields can 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: Although this work was reviewed by NOAA and EPA and approved for publication, it may not reflect official Agency policy.

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.
- —, T. Pierce, A. J. Cimorelli, J. A. Herwehe, T. Palma, and R. Tang, 2004: Linking air toxic concentrations from CMAQ to the HAPEM-5 exposure model at neighborhood scales for the Philadelphia area. From the 16th Conference on Biometeorology and Aerobiology, Vancouver, BC, Canada, Amer. Meteor. Soc.
- 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 <u>http://www.itl.nist.gov/div898/handbook/</u>.]