2.04 Temporal Collinearity Amongst Modeled and Measured Pollutant Concentrations and Meteorology

V.Garcia[1], P.S. Porter[2], E.Gégo[3], and S.T. Rao[4],
[1]Atmospheric Modeling Division, U.S. Environmental Protection Agency, garcia.Val@epamail.epa.gov
[2]University of Idaho, Idaho Falls, ID, porter@if.uidaho.edu
[3]Porter-Gego and Associates, e.gego@onewest.net
[4]Atmospheric Modeling Division, U.S. Environmental Protection Agency, rao.st@epa.gov,

Key topic: 7 - Air quality effects on human health, ecosystems and economy

The results from epidemiology time series models that relate air quality to human health are often used in determining the need for emission controls in the United States. These epidemiology models, however, can be sensitive to collinearity among co-variates, potentially magnifying biases in the parameter estimates caused by exposure misclassification error or other deficiencies in the time series models by orders of magnitude. As a result, we examined collinearity among several covariates typically used in air quality epidemiology time series studies (ozone, fine particulate matter and its species, and temperature). In addition, we examined the ability of a bias-correction technique applied to estimates simulated by the Community Multiscale Air Quality (CMAQ) model to "fill-in" for the spatial and temporal limitations of observations for purposes of reducing exposure misclassification. Specifically, we evaluated whether the bias-adjusted CMAQ estimates could replicate the correlation among variables seen in the observations. The results presented are for a domain east of the Rocky Mountains for the entire 2006 year and indicate that collinearity among covariates varies across space.

2.04.1 Introduction

The United States Environmental Protection Agency (USEPA) relies predominantly on epidemiology time series studies to estimate future health impacts of emission controls [1]. High correlation among explanatory variables used in these studies can result in inaccurate results when applied in health impact assessments. In addition, the assignment of the wrong exposure value differentially (misclassification) can mask the true health effect of a pollutant [2]. To address these two issues, this study examines (1) the collinearity that exists among ozone, particulate matter and its species, and temperature; and (2) whether we can fully represent this natural relationship among covariates in the deterministic, 3-dimensional Community Multiscale Air Quality (CMAQ) model. This latter objective is particularly relevant due to the paucity of measurements for some pollutants (e.g., speciated fine particulate matter) that are typically measured once- in-3-days at relatively few locations in the U.S.

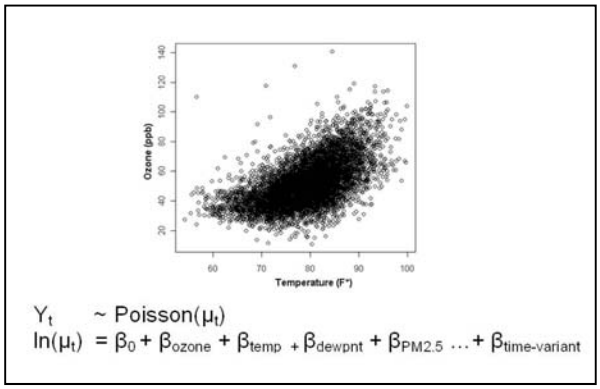

Figure 2.04.1 Example of variables used in an epidemiology Poisson regression model. Scatter plot shows relationship between ozone and temperature ($R^2 = .58$).

Figure 2.04.1 shows an example of a typical time series model used in an epidemiology study and the collinearity between ozone and temperature for 10 summers (1997-2006) in New York State. Collinearity can also exist between ozone and fine particulate matter. Multicollinearity among covariates can magnify the effect of bias introduced by autocorrelation, misclassification error or other model deficiencies by orders of magnitude. Thus, while it doesn't necessarily reduce the predictive power or reliability of a model when all collinear covariates are included, it can affect calculations regarding individual predictors, such as when the main effects coefficients are used to assess health impacts. If the covariates (and relationship among them) in the new dataset differ from the data that was fitted, it can introduce large errors in predictions [3].

Misclassification (inaccurately assigning exposure across space or time differentially) can also introduce errors in epidemiology time series studies. Air quality models, such as CMAQ, can help fill in missing measurement data but contains some bias due to uncertain emissions and meteorology input data, as well as limited knowledge of the physical and chemical processes governing the formation of ambient pollutants. Hence, CMAQ model estimates were combined with observations (Garcia et al., 2010) to produce bias-adjusted pollutant estimates with the objective of providing more spatially and temporally complete ambient air pollutant data for use in epidemiology studies. As part of this study, we examined the ability of the model alone, and the bias-adjusted model to reproduce the relationships that exist in the observed covariates used in a standard epidemiology study.

The objectives of this study were to (1) examine whether the model (CMAQ and "adjusted" CMAQ) are capturing the temporal variability seen in observations, (2) examine whether the relationships among pollutants and temperature estimated by CMAQ and adjusted CMAQ reflect those seen in observations and (3) understand what pollutants are correlated with each other and with temperature.

Approach

Maximum daily 8-hr averaged ozone (O_3), and 24-hr averaged fine particulate matter ($PM_{2.5}$), sulfate (SO_4), nitrate (NO_3), ammonium (NH_4), elemental carbon (EC) and organic carbon (OC) were calculated from measurements obtained from the USEPA's Air Quality System database (http://www.epa.gov/oar/data/aqsdb.html) for 2006. Measurements of O_3 and $PM_{2.5}$ were available for each day, whereas, measurements for $PM_{2.5}$ species were available for 1-in-3 days only. Daily averages were also calculated from the hourly concentrations simulated by the CMAQ model v.4.5 at a 12 km horizontal grid resolution. The meteorology and emissions inputs for this simulation were from the Fifth-Generation NCAR / Penn State Mesoscale Model (MM5) and EPA's 2001 National Emissions Inventory, respectively. The 12-km simulation encompassed most of the Eastern U.S. and was nested within a 36 km x 36 km horizontal grid simulation covering the contiguous U.S. using the same model configuration as the 12-km nested simulation. Finally, the observations and modeled estimates were combined using a multiplicative adjusted bias approach described in [4] to produce daily averaged estimates as described above. To summarize the process, the ratio of observed to modeled values was calculated for each grid cell containing an observation. These ratios were interpolated using a kriging technique and then applied to the CMAQ estimates to produce a bias-adjusted value for each CMAQ grid cell. Pearson correlation coefficient (R) was calculated to measure the temporal dependencies.

Discussion and Results

Bias-adjusted CMAQ estimates captured the temporal variability seen in observations for most pollutants (objective 1). Challenges, however, still remain in estimating EC because of the high spatial and temporal heterogeneity of this pollutant. In addition, spatial differences in capturing the observed temporal variability were seen for OC along the Southern coastline and NO_3 along the Western edge of the domain and in the South. With regard to collinearity among variables (objectives 2 and 3), ozone is positively correlated with $PM_{2.5}$, SO_4 and temperature at most sites, reflecting the dominant SO_4 component of $PM_{2.5}$ mass and its common source with secondarily formed ozone from photolysis. Strong spatial patterns existed for several pollutants, with very strong spatial correlations between ozone and nitrate and ammonium in the Southeastern U.S. (not shown). As expected, $PM_{2.5}$ is highly correlated with most of its constituents (Figure 2.04.2), but surprisingly, not as correlated with NO_3, perhaps due to seasonal differences (e.g., high correlation in winter, but relatively low correlation in summer for the domain studied). The correlation between $PM_{2.5}$, and EC and OC is strongest in the upper Northwest portion of the domain, most likely due to wood burning. Correlation between $PM_{2.5}$ and NH_4 is dominant in the Eastern U.S.

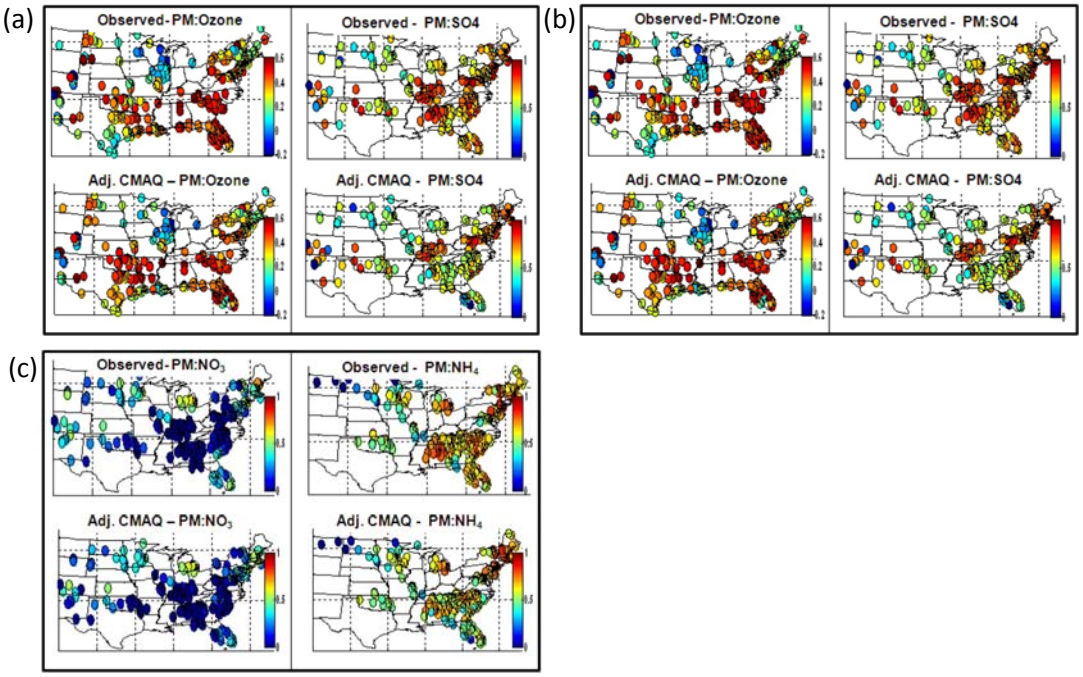

Figure 2.04.2: Collinearity between PM2.5, and ozone and SO4 (panel a), EC and OC (panel b), and NO3 and NH4 (panel c). Second row of each panel shows ability of bias-adjusted CMAQ to replicate the collinearity among observed pollutant concentrations shown in first row. Each circle represents Pearson's correlation (R) between the indicated pollutants across 365 days at each monitoring location.

Disclaimer: The United States Environmental Protection Agency through its Office of Research and Development funded and collaborated in the research described here under EP-D-10-078 to Porter-Gego. It has been subjected to Agency review and approved for publication.

References

1. US Environmental Protection Agency (2010) Our Nation's Air Status and Trends through 2008. Report No. EPA-454/R-09-002.

2. Rothman K J (2002) Epidemiology: an introduction. New York City: Oxford University Press.

3. Farrar D E & Glauber RR (1967) Multicollinearity in regression analysis: the problem revisited. Rev Econ Stat 49:92-107.

4. Garcia VC, Foley KL, Gego E, Holland DM, Rao ST (2010) A Comparison of Statistical Techniques for Combining Modeled and Observed Concentrations to Create High-Resolution Ozone Air Quality Surfaces. J. Air Waste Manage. Assoc. 60:586 - 595.