Integration of air quality modeling and monitoring data for enhanced health exposure assessment

For the special edition of EM Magazine "Monitoring and Modeling Needs in the 21st Century."

Bruce Denby[1*], Val Garcia[2], David Holland[2] and Christian Hogrefe[3,4]

[1] The Norwegian Institute for Air Research (NILU)

[2] U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory

[3] Bureau of Air Quality Analysis and Research, New York State Department of Environmental Conservation

[4] Atmospheric Sciences Research Center, University at Albany

*** Corresponding author**

Bruce Denby (bde@nilu.no)

The Norwegian Institute for Air Research (NILU)

PO BOX 100

2027 Kjeller, Norway

Tel: +47 63898164

In order to assess the environmental impact of air pollution on human health it is necessary to establish the concentrations to which the population is exposed. The obvious way to determine this is to measure these quantities. However, given the limited number of monitoring stations available, how is it possible to provide spatially distributed pollution concentrations far from monitoring sites in order to assess the exposure of an entire population?

Traditionally ground based monitoring has been used to provide air quality information since it is expected to give the best estimate. This may be suitable when a very limited area is to be assessed, e.g. in occupational health studies, or when monitoring data is representative of a large area, e.g. in rural regions, but generally such monitoring has a limited spatial

representativeness. This can be problematic in urban areas since there can be significant variation in air quality due to the heterogeneity of the emissions sources and the complex flow patterns caused by urban morphology.

Several studies have demonstrated that an accurate assessment of temporal and spatial variations in ambient concentrations is critical for the interpretation of time-series epidemiology studies. Health studies have shown (e.g. Sarnat et al. 2006 and Jerrett et al. 2005) that a more narrow definition of the geographic domain of the study populations leads to stronger associations between exposure and health outcomes (e.g., hospital admissions, mortality counts). In order to improve epidemiology and health impact studies enhanced spatial and temporal coverage and resolution is thus required.

The immediate solution is to apply spatial interpolation techniques to the available monitoring data to provide air quality information between monitoring stations. Such interpolation methods may be geometric in nature, e.g. linear interpolation or inverse distance weighting, or they may be statistically based, such as kriging interpolation methods. No matter the interpolation method applied, the amount and density of the available monitoring data is usually limited and interpolation alone cannot provide information concerning the spatial variability of the concentrations between the measurement sites.

To improve the spatial representativeness of the monitoring data it is necessary to make use of other related supplementary data sources that have a better spatial coverage than the monitoring data itself. Such supplementary information may include distances from major roads, traffic volumes, population density, land use characteristics, satellite data, etc.. Though it is possible to use these data directly through a range of spatial statistical methods, it is the **air quality model** that best describes the relevant physical and chemical processes and provides high spatial and temporal resolution data that can be used for improving the coverage of the monitoring information. The major drawback of modeling is it's level of uncertainty, which is usually significantly higher than that for monitoring. It is therefore advantageous to combine the monitoring and modeling data sources in an optimal way to produce spatio-temporal maps of the pollutants.

What is interpolation, data fusion, data integration and data assimilation?

There are a number of terms used to describe the combination of different data sources. 'Interpolation' refers to methods that use monitoring as the primary dataset and, based on these data and possibly other supplementary data, predict concentrations at any arbitrary point in space (e.g. Beelen, 2009). Methods that combine various data sources, without directly considering one or the other to necessarily be primary, are often referred to as 'data fusion' or 'data integration' methods. They take any number of datasets and combine these in a range of ways, either through geometric means or based on statistical optimization methods. For example, it is possible to fuse interpolated monitoring data, satellite data and air quality modeling data into a single integrated map (e.g. Sarigiannis et al., 2004). The fusing will most likely take the form of a weighted linear combination of the different data sources, with the weighting being dependent on the estimated uncertainty of each of the data sources. Data fusion and interpolation methods are generally not concerned with any physical or chemical constraints but are mainly subject to statistical constraints.

'Data assimilation' refers to a modeling technique that incorporates monitoring data directly into air quality model calculations during the modeling process itself. It is the measured data that helps guide the model towards an optimal solution, and one that is consistent with the physical description provided by the air quality model. The most common type of data assimilation applied are the variational methods (Elbern et al., 1999), which are also extensively used in meteorological forecast, but other methods such as Ensemble Kalman filters (van Loon et al., 2000) may also been applied. Data assimilation is now used operationally in air quality forecasting (e.g. Sahu et al., 2009) and it is also applied for air quality assessment purposes (Denby et al, 2008). Data assimilation is most often applied on the regional scale and is rarely applied on the urban scale, due to the complexities of the urban environment. As a result it is less applicable for health applications in urban regions.

Examples of mapping methods using monitoring and air quality modeling

There are thus a number of methods available that can be applied to combine monitoring and modeling data. These range from simple statistical methods to complex data assimilation models. One of the most straightforward methods is multiple linear regression, where model concentrations, and other supplementary data, are fitted to the available observations using least squares optimization (e.g. Horálek et al., 2007; Denby and Pochmann, 2007), see figure 1. Though this will provide an unbiased model field there may still be significant deviations from the observations. This deviation may be accounted for by using residual interpolation of the

deviations. In this way the model field provides the basis for the concentration map and the residual deviations are accounted for by using interpolation methods (e.g. Horálek et al., 2007; Kassteele et al., 2007; Hogrefe et al., 2009). An example of this method applied to all of Europe at a resolution of 10 km is presented in figure 2.

There are also a number of more complex statistically based methods for achieving data fusion. Such methods include those described by Fuentes and Raftery (2005), Gelfand and Sahu (2009) and McMillan et al. (2009), figure 3. These methods combine Bayesian approaches with a range of statistical methods. A good example of the potential of data fusion methods is that described by van de Kassteelle et al. (2006) where satellite remote sensed data, ground based monitoring data and meso-scale air quality modeling data have been combined to provide annual mean concentrations of PM_{10} for all of Europe.

Future directions

There is an increased activity in research aimed at data fusion and data assimilation, particularly in regard to air quality forecasting, but also for improved exposure assessment. Future epidemiological and exposure studies will be making more and more use of the air quality model and its enhanced spatial resolution. Even now many studies use concentrations at home addresses based on modeling, rather than monitoring.

Other researchers are also now beginning to use air quality dispersion models combined with micro-environmental personal exposure modeling tools to support air pollution exposure and health studies. The advantage of combining air quality and exposure models is that they can take account of exposure to indoor and outdoor sources, in the same manner that the personal monitoring data can (e.g., Georgopoulos, 2005; Isakov, 2009). Such methods require as accurate as possible description of the spatial and temporal resolved concentration fields, something that the data fusion methods aim to provide.

There are still a large number of challenges in optimally combining the various datasets and applying these to health studies. These include matching the spatial representativeness of the different data sources in a suitable way, designing monitoring networks for data fusion purposes, improving estimates of the uncertainties for the optimal combination of the datasets, improving spatial resolution and improving the links to exposure modeling. These tasks will involve the coming together of a multiple of disciplines, requiring that air quality modelers and

monitors, statisticians and exposure and health modelers share a common goal and speak a common language.

Disclaimer

The U.S. Environmental Protection Agency's Office of Research and Development partially collaborated in the research described here. Although it has been reviewed by EPA and approved for publication, it does not necessarily reflect the Agency's policies or views.

References

Beelen R., G. Hoek, E. Pebesma, D. Vienneau, K. de Hoogh, D. J. Briggs (2009) Mapping of background air pollution at a fine spatial scale across the European Union Science of The Total Environment, Volume 407, Issue 6, Pages 1852-1867.

Denby, B. and Pochmann, M. (2007). Basic data assimilation methods for use in urban air quality assessment. In: Proceedings of the 6th International Conference on Urban Air Quality. Limassol, Cyprus, 27-29 March 2007. Ed. by R.S. Sokhi and M. Neophytou. Hatfield, University of Hertfordshire(CD-ROM).

Denby B., M. Schaap, A. Segers, P. Builtjes and J. Horálek (2008). Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale. Atmos. Environ. 42, 7122-7134.

Elbern, H. and H. Schmidt (1999): A four-dimensional variational chemistry data assimilation scheme for eulerian chemistry transport modeling. JGR, 104, 18583-18598.

Fuentes, M., and Rafter, A. (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. Biometrics 61, 36-45.

Gelfand, A. E., and Sahu, S. K. (2008). Combining Monitoring Data and Computer Model Output in Assessing Environmental Exposure. In: The Handbook of Bayesian Analysis. (O'Hagan A,West M, eds). Oxford: Oxford University Press.

Georgopoulos, P. G., S.-W. Wang, et al. (2005). "A source-to-dose assessment of population exposures to fine PM and ozone in Philadelphia, PA, during a summer 1999 episode." J Expo Anal Environ Epidemiol **15**(5): 439-457.

Hogrefe, C., B. Lynn, R. Goldberg, C. Rosenzweig, E. Zalewsky, W. Hao, P. Doraiswamy, K. Civerolo, J.Y. Ku, G. Sistla, and P.L. Kinney (2009) Combined Model-Observation Approach to Estimate Historic Gridded Fields of PM2.5 Mass and Species Concentrations. Atmospheric Environment. Volume 43, 2561-2570

Horálek J., P. Kurfürst, P. de Smet, F. de Leeuw, R. Swart, T. van Noije, B. Denby and J. Fiala (2007) Spatial mapping of air quality for European scale assessment. ETC/ACC Technical Paper 2006/6. http://air-climate.eionet.europa.eu/reports/ETCACC_TechnPaper_2006_6_Spat_AQ

Isakov, V., Touma, J., Burke, J., Lobdell, D., Palma, T., Rosenbaum, A., Özkaynak, H. (2009). Combining Regional and Local Scale Air Quality Models with Exposure Models for Use in Environmental Health Studies. J. A&WMA. 59: 461-472.

Jerrett M., Burnett R.T., Ma R.J., Pope C.A., Krewski D., Newbold K.B., Thurston G., Shi Y.L., Finkelstein N., Calle E.E., and Thun M.J. (2005). Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology: 16(6): 727-736.

Kassteele, van de, J., A. Stein, A. L. M. Dekkers and G. J. M. Velders (2007). External drift kriging of NOx concentrations with dispersion model output in a reduced air quality monitoring network. Environmental and Ecological Statistics, online publication. DOI: 10.1007/s10651-007-0052-x.

Kassteele, van de J., R. B. A. Koelemeijer, A. L. M. Dekkers, M. Schaap, C. D. Homan and A. Stein (2006) Statistical mapping of PM10 concentrations over Western Europe using secondary information from dispersion modelling and MODIS satellite observations. Stoch Environ Res Risk Assess, Vol 21, pp. 183–194

McMillan, N., Holland, D. M., Morara, M., and Feng, J. (2009). Combining numerical model output and particulate data using Bayesian space-time modeling. Accepted in Environmetrics.

Sahu, S., Yip, S., and Holland, D. M. (2009). Improved space-time forecasting of next day ozone concentrations in the eastern U.S.. Atmospheric Environment, 43, 494–501.

Sarigiannis, D. A., N. A. Soulakellis, and N. I. Sifakis (2004). Information Fusion for Computational Assessment of Air Quality and Health Effects. Photogrammetric Engineering & Remote Sensing Vol. 70, No. 2, pp. 235–245.

Sarnat S.E., H. H. Suh, B. A. Coull, J. Schwartz, P. H. Stone and D. R. Gold (2006). Ambient particulate air pollution and cardiac arrhythmia in a panel of older adults in Steubenville, Ohio *Occup Environ Med.* 63:700-706. doi:10.1136/oem.2006.027292

van Loon, M, P.J.H. Builtjes, A. Segers, 2000. Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. Environmental Modeling and Software 15, 603-609.

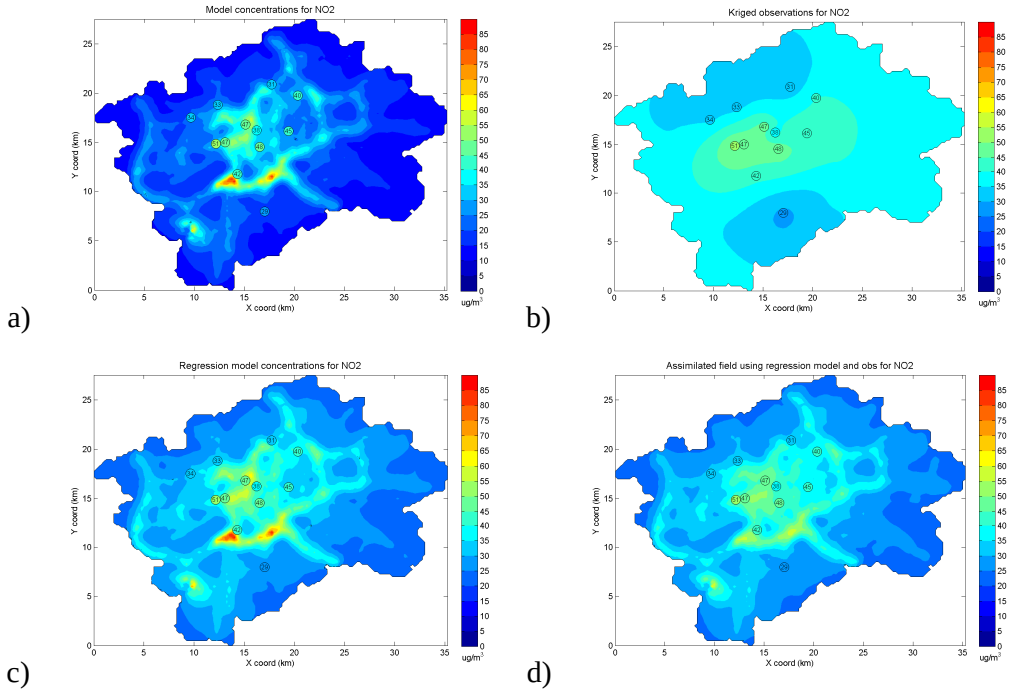

Figure 1. Application of different statistical interpolation methods for Prague, annual mean NO_2, using modelling data with a resolution of 250 m and 11 monitoring sites. a) Modelled concentrations and observations (numbered circles). b) Ordinary kriging of the observations. c) Model fields after regression with observations. d) Weighted combination of the fields b) and c) using a Bayesian approach (Denby and Pochmann, 2007).

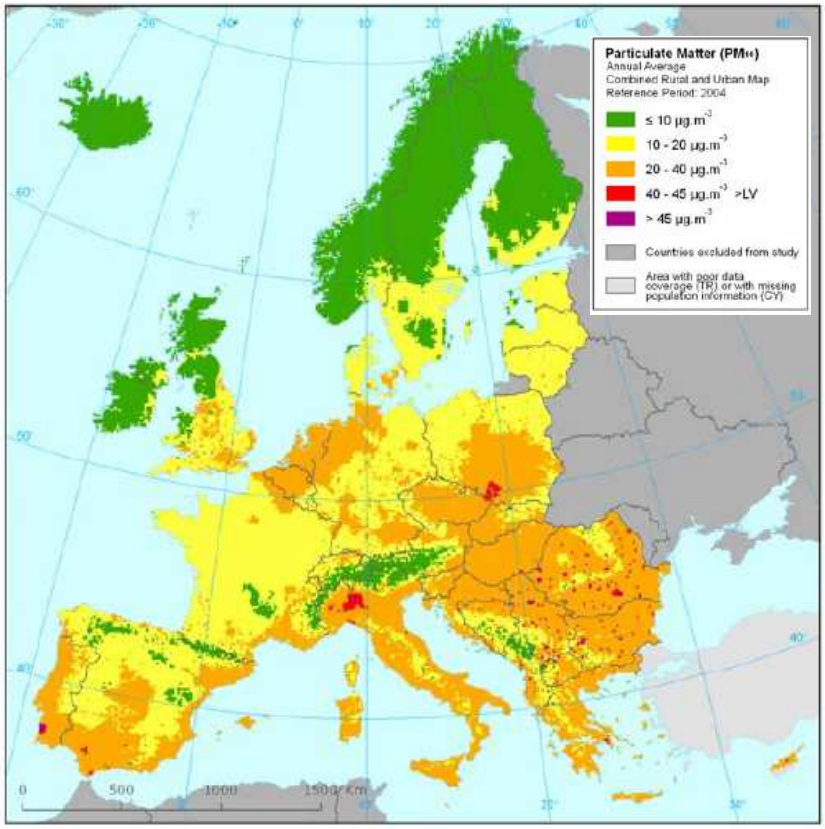

Figure 2. Annual mean PM_{10} maps generated using multiple linear regression and residual kriging using topography, meteorology and air quality modelling data. Resolution of the map is 10 km and the reference year is 2005 (Horálek et al. 2007).

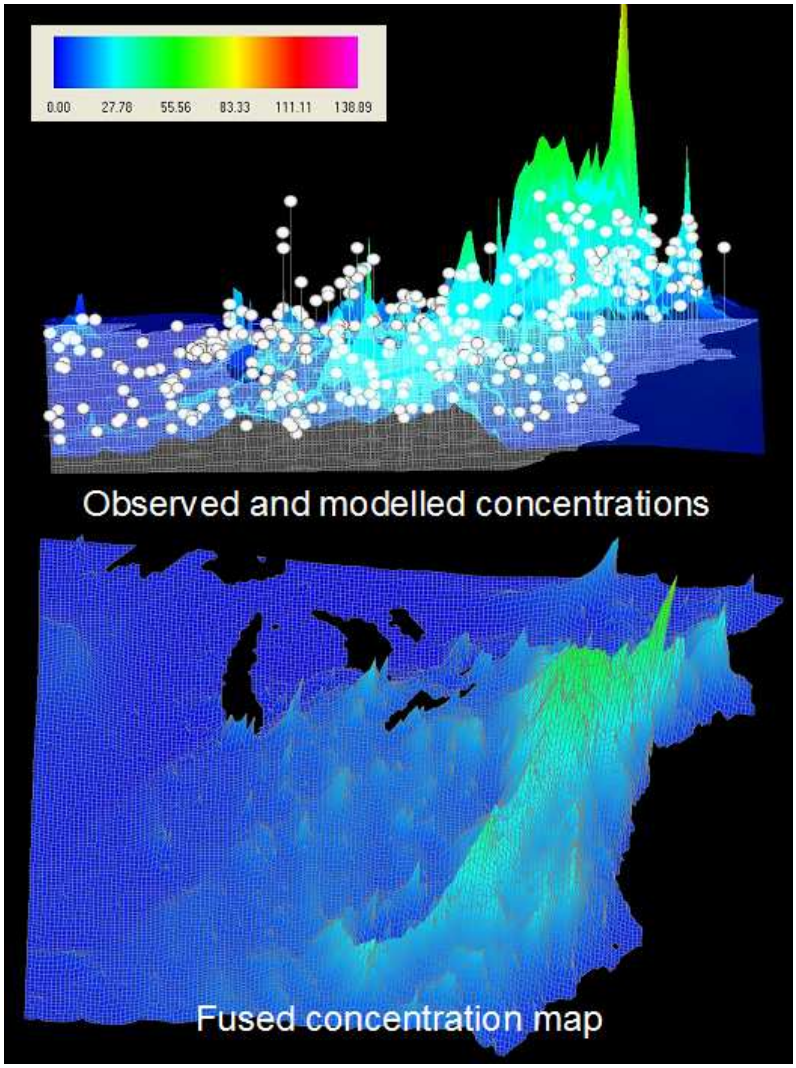

Figure 3. Example of data fusion using a hierarchical Bayesian technique (McMillan et al. 2009) showing fine particulate matter concentrations (μgm^{-3}) for February 9, 2001: Top shows the model simulation (underlying surface) overlaid with observations (white circles) prior to the data fusion. Bottom is the combined surface map.