Presented at the 6th Annual CMAS Conference, Chapel Hill, NC, October 1-3, 2007

STATISTICAL METHODS FOR MODEL EVALUATION - MOVING BEYOND THE COMPARISON OF MATCHED OBSERVATIONS AND OUTPUT FOR MODEL GRID CELLS

Jenise L. Swall and Kristen M. Foley*
Atmospheric Sciences Modeling Division, Air Resources Laboratory, National Oceanic and Atmospheric Administration, RTP, NC, 27711, USA
(In partnership with the U.S. Environmental Protection Agency, National Exposure Research Laboratory)

1. INTRODUCTION

Standard evaluations of air quality models rely heavily on a direct comparison of monitoring data matched with the model output for the grid cell containing the monitor's location (e.g. Eder and Yu 2006, Appel et al. 2007). While such techniques may be adequate for some applications, conclusions are limited by such factors as the sparseness of the available observations (limiting the number of grid cells at which the model can be evaluated), potential measurement error in the observations, and the incommensurability between volume-averages and point-referenced observations. While we focus most closely on the latter problem, we find that it cannot be addressed without some discussion of the others. Our approach uses simulated datasets to demonstrate cases in which incommensurability is more likely to adversely affect a traditional analysis. Future work will illustrate the impact on model evaluation analysis using a comparison of CMAQ simulations and observed maximum 8 hour ozone.

2. SIMULATED "PERFECT-WORLD" EXAMPLE WITH WEAK CORRELATION

For the purposes of illustration, we simulate pollutant fields with both strong and weak correlation structures. Fig. 1 shows an example of such a field with weak spatial dependence. The superimposed gray lines show where our "grid cell" boundaries are located. The circles designate 26 points chosen to represent observations in this domain. Using this field, we can find the average in each grid cell, as shown in Fig. 2.

We can now compare our hypothetical observations with the averages of the cells in which they fall. This allows us to assess the outcome of a traditional analysis technique assuming a perfect scenario in which the model output is in perfect agreement with the "true" process and there is no measurement-related or fine-scale variability in the observations. The scatterplot in Fig. 3 shows the observations vs. the grid cell averages, with a red 1:1 line shown for reference. We note that the correlation between the two is only about 0.84. This case study confirms that in a situation in which the extent of the spatial correlation is quite short-range, a traditional analysis can yield misleading results, even if the model and observations are actually in perfect agreement. However, this situation becomes much less extreme when the spatial correlation is farther-reaching, i.e. when the field is more spatially homogeneous.

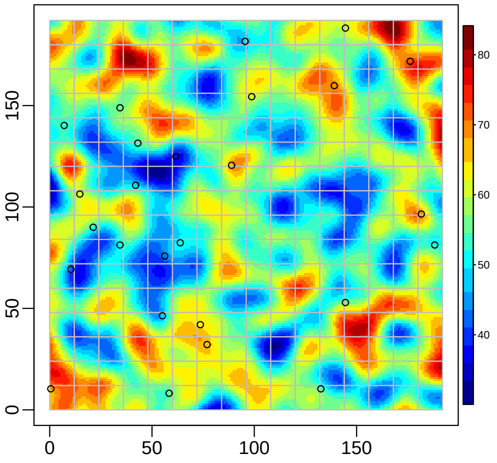

Fig. 1. Simulated data with short-range spatial dependence. In this case, the correlation between neighboring sites becomes negligible after approximately 30 units.

**Corresponding author:* Kristen M. Foley, U.S. EPA, MD E243-01, RTP, NC 27711; e-mail:foley.kristen@epa.gov

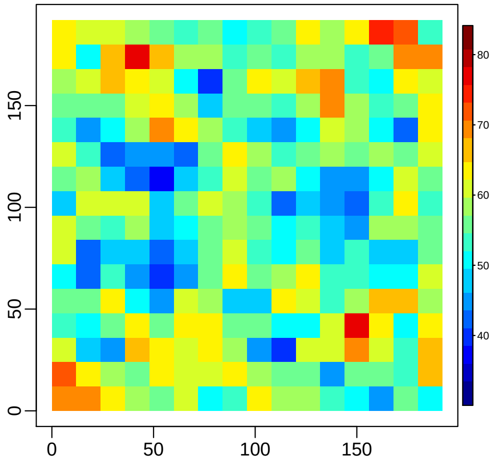

Fig. 2. Grid cell averages based on the simulated data portrayed in Fig. 1. Each grid cell is a square with a side length of 12 units.

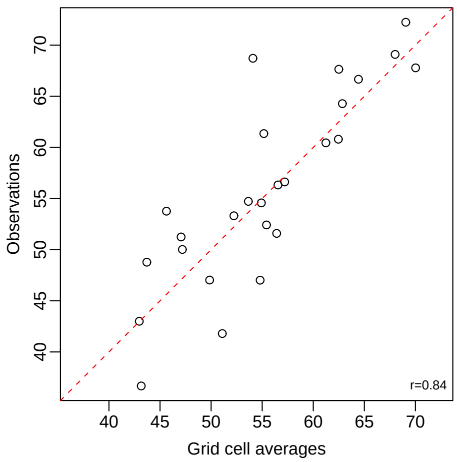

Fig. 3. Observations vs. grid cell averages, based on the data shown in Fig. 1 and Fig. 2.

3. SIMULATED EXAMPLE WITH STRONGER CORRELATION AND MEASUREMENT ERROR

Fig. 4 gives an example of simulated field with correlation that becomes effectively zero only at distances greater than about 160 units. This means that points near the center of our region are still correlated with those at the edges, at least to some extent, accounting for the smooth, homogeneous nature of the picture. Fig. 5 shows the grid cell averages calculated based on this field.

Unlike the example in the previous section, in this case we allow for a slightly more realistic situation in which the observations, while taken at the locations designated by the black circles in Fig. 4, have an additional error component to represent measurement or other fine-scale error. Note that we are still calculating the grid cell averages directly from the underlying field shown in Fig. 4, so we continue to assume that the model output is equal to this “true” simulated field, i.e. no model error.

Fig. 6 shows the observations (with error as described above) plotted against the averages for grid cells in which they fall. As before, we see substantial variability around the red 1:1 reference line, with a correlation coefficient of approximately 0.87. However, further analysis shows that in this more spatially homogeneous field, most of this variability comes from the error associated with the observations, rather than from the incommensurability issue.

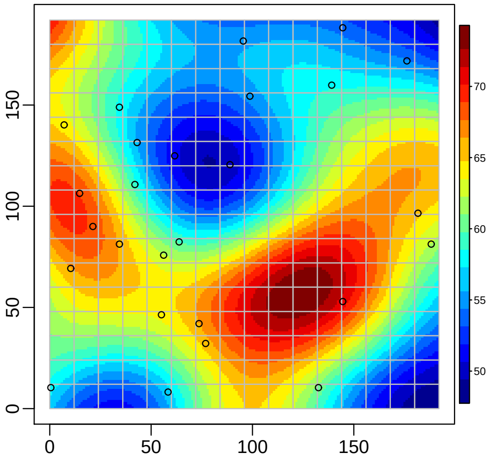

Fig. 4. Simulated data with long-range spatial correlation structure.

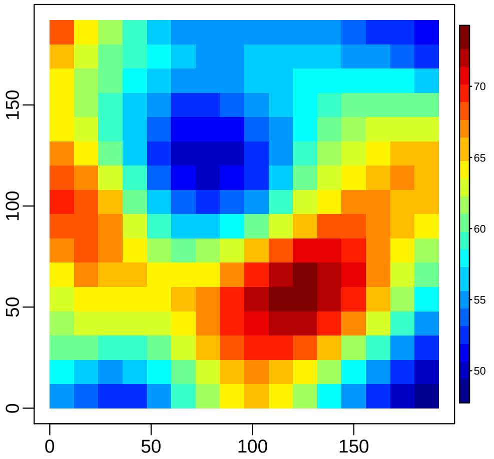

Fig. 5. Grid cell averages based on the simulated data portrayed in Fig.4. Each grid cell is a square with a side length of 12 units.

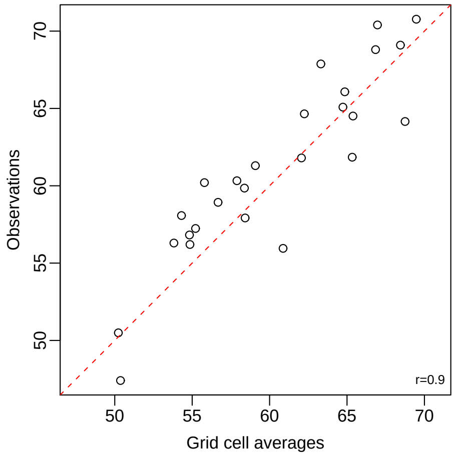

Fig. 6. Observations vs. grid cell averages, based on the data shown in Fig. 4 and Fig. 5.

4. SIMULATED EXAMPLE COMPARING KRIGING TECHNIQUES FOR USE IN MODEL ASSESSMENT

In the preceding sections, the comparison of simulated grid cell averages and observations was facilitated mostly by the scatterplots in Fig. 3 and Fig. 6. Using this approach, only grid cells which contain observations can be evaluated. As mentioned in the introduction, we are interested in investigating statistical techniques which make use of the observational data to estimate the level of the field at unobserved locations. We can then compare the grid cell averages with these estimates.

Given knowledge of or an estimate of the spatial correlation structure, kriging is a technique which can be used to make such estimates. While we are interested in other techniques as well, we focus on kriging here primarily because of its ease of use (it is available in many software packages) and its ability to provide error estimates associated with its predictions. We return to our simulations and the discussion in the preceding sections in an attempt to illustrate how incommensurability affects our choice of kriging techniques.

4.1 Kriging to the grid cell centers

A typical kriging procedure begins by assessing the correlation structure inherent in the spatial field. In our case, since the data are simulated, this correlation structure is known. In more realistic practice, this would rarely be the case, so techniques such as variogram estimation might be used to assess this structure.

The practitioner must identify the locations at which estimates are desired. In classical kriging, this would just be a series of points, most probably the centers of the grid cells to be evaluated. The kriging procedure provides the estimate at each of these points and an estimate of the standard error, based on the provided observational and correlation information.

Fig. 8 shows the kriging predictions made based on the observations shown in Fig. 7. Comparison with Fig.5, which contains the true grid cell averages for this simulation, shows that our estimates are reasonable ones for the most part. As is the case with most such estimation procedures, we see notable discrepancies where observational information is most limited, such as at the edges and in the right portion of the region. Fig. 8 also shows the well-known smoothing effect of kriging (and other similar methods), in which the extremes are not always well-captured. This is particularly noticeable in the “hot spot” in the right portion of the region, where the situation is further complicated by the relative scarcity of observational data.

We can also judge the performance of the kriging technique based on Fig. 9, which shows (for a subset of the grid cells) each kriging estimate vs. the true grid cell value (represented by points). The plot also displays 95% confidence bounds, determined using the kriging error estimates. In this case, the intervals based on kriging to the center points of all 256 grid cells (not just the subset shown), include the true value. We would expect approximately 243 (95% of 256) of the intervals to miss their target. This indicates that the error estimates are not well-calibrated for estimating spatial average, and we explore this issue further in the next section.

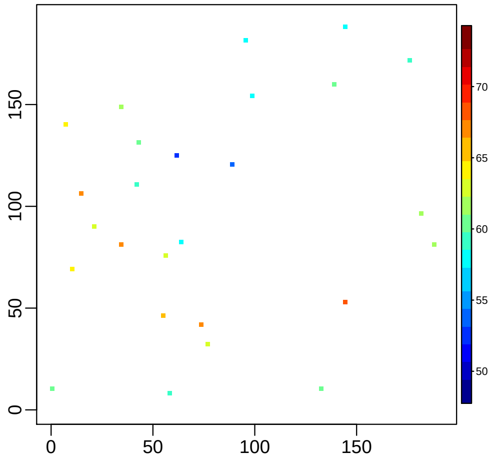

Fig. 7. Observations simulated with fine-scale error at the locations shown in Fig. 4 (long-range spatial correlation)

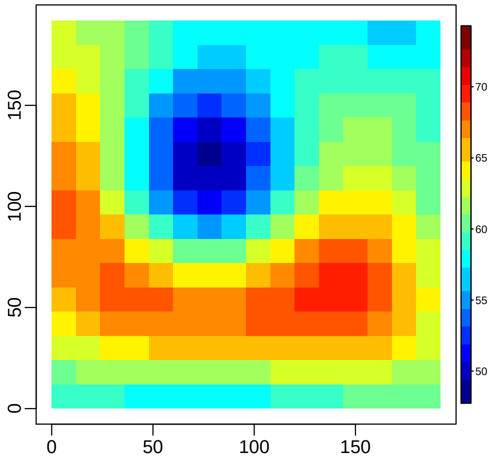

Fig. 8. Estimates yielded by kriging the observations in Fig. 7 to the grid cell centers (long-range spatial correlation)

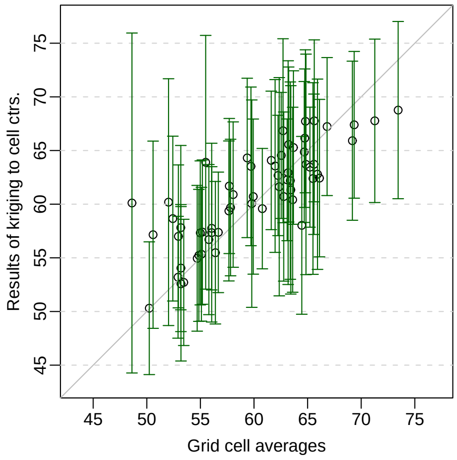

Fig. 9. Sample of results yielded by kriging to the grid cell centers (long-range spatial correlation)

4.2 Block kriging

Given our discussion of the potential problems associated with incommensurability and the calibration issues discussed in the previous section, it is reasonable to question whether we should be using the kriging technique to estimate the spatial field point-wise at the grid cell centers. Statistical reasoning (Gelfand et al. 2001) indicates that estimates of variability and error are generally sensitive to the incommensurability issue, with estimates of averages less variable than estimates for individual points. The block kriging technique (e.g., Goovaerts 1997, pg. 152) allows us to adjust for incommensurability by estimating the spatial field at a lattice of points within each of the grid cells. The estimate for the grid cell average is then given by the average of the estimates for all the lattice points, and the variance of the estimate is given by a function of the covariances among all the lattice points.

Fig. 10 shows the block kriging estimates for each grid cell, based once again on the observational data in Fig. 7. Comparison with Fig. 8 shows that the estimates given by the two kriging techniques are very similar, with only minor differences visible on close inspection. This is what we would expect, especially with the effective correlation range extending well beyond the distance spanned by not just one, but several, grid cells.

However, Fig. 11 shows that the error estimates and resulting confidence intervals (in green if the true value is captured and red otherwise) are quite different for the two kriging techniques. The confidence intervals depicted in Fig. 11 are noticeably shorter than those in Fig. 9, due to smaller errors associated with estimating an average rather than an individual location. The block kriging method showed better calibration, with 247 (about 96%) of the confidence intervals enclosing the actual grid cell average. This further indicates that the errors associated with kriging to the cell centers are estimated to be higher than what is needed for kriging to grid cell averages. Achieving a more accurate estimate of the error associated with the kriging technique allows us to better assess model performance by giving us information about differences which are more likely due to estimation error than to model error.

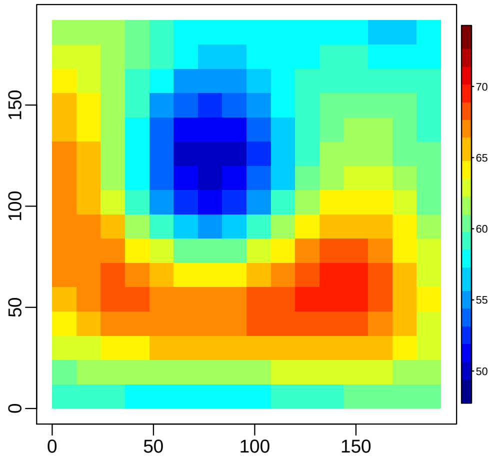

Fig. 10. Estimates yielded by block kriging the observations in Fig. 7 (long-range spatial correlation)

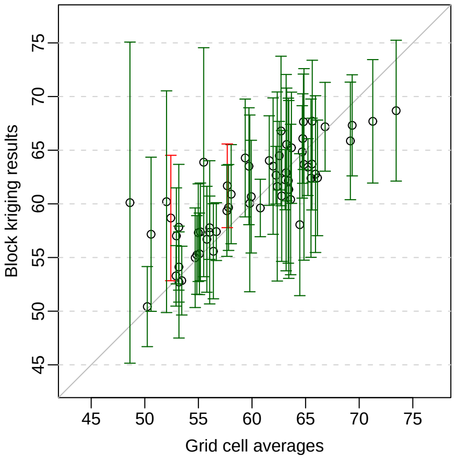

Fig. 11. Sample of results yielded by block kriging (long-range spatial correlation)

4. DISCUSSION

Analysis such as block kriging provides one approach to address the difference in variability between a point measurement and a volume-average prediction by using a statistical model to characterize sub-grid variability based on observed values. An alternative approach is to use a nested grid modeling system of fine scale features to simulate sub-grid variability as proposed by Ching et al. (2006). Although in the context of air quality modeling most evaluations have not considered this issue of incommensurability in detail, some authors have considered the problem and developed alternative methods for comparing measurements and model output (Swall and Davis 2006). The more sophisticated statistical modeling provides additional information, which is not available in a matched model to observation type comparisons. Spatial analysis of model errors is used to determine regions where model output is significantly different from observation-based estimates. These areas may be used for diagnostic evaluation to identify the source of consistent model errors. The added benefit of this extra layer of analysis will depend on the goals of a particular model evaluation. Analysis of observed and modeled ozone data will be used to further compare standard evaluation methods and more complex statistical modeling in an operational setting.

DISCLAIMER: The research presented here was performed under the Memorandum of Understanding between the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under agreement number DW13921548. This work constitutes a contribution to the NOAA Air Quality Program. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views.

References

Appel, K. W., Gilliland, A. B., Sarwar, G. and R. C. Gilliam, 2007: Evaluation of the Coummunity Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part I – ozone, *Atmos. Envir., to appear.*

Ching, J., Herwehe, J., and J. Swall, 2006: On joint deterministic grid modeling and sub-grid variability conceptual framework for model evaluation, *Atmos. Envir.*, **40**, 4935-4945.

Eder, B., and S. Yu, 2006: A performance evaluation of the 2004 release of Models-3 CMAQ. *Atmos. Envir*., **40**, 4894-4905.

Gelfand, A. E., Zhu, L., and B. P. Carlin, 2001: On the change of support problem for spatio-temporal data, *Biostatistics*, **2,** 31-45.

Goovaerts, P. 1997: *Geostatistics for Natural Resources Evaluation*. Oxford University Press, 483 pp.

Swall, J. and J. Davis, 2006: A Bayesian Statistical Approach for Evaluation of CMAQ, *Atmos. Envir.*, **40**, 4883-4893.