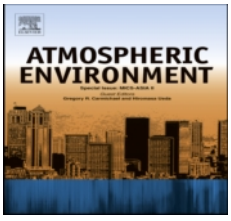


Title: The impact of spatial correlation and incommensurability on model evaluation

Authors: Jenise L. Swall

PII: S1352-2310(08)00993-X
DOI: 10.1016/j.atmosenv.2008.10.057
Reference: AEA 8679

To appear in: *Atmospheric Environment*

Received Date: 30 July 2008
Revised Date: 16 October 2008
Accepted Date: 19 October 2008

Please cite this article as: Swall JL,. The impact of spatial correlation and incommensurability on model evaluation, Atmospheric Environment (2008), doi: 10.1016/j.atmosenv.2008.10.057

The impact of spatial correlation and incommensurability on model evaluation

Jenise L. Swall*, Kristen M. Foley

Atmospheric Modeling Division, National Exposure Research Laboratory, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

Abstract

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. 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) and the incommensurability between volume-averages and pointwise observations. We examine several sets of simulations to illustrate the effect of incommensurability in a variety of cases distinguished by the type and extent of spatial correlation present. Block kriging, a statistical method which can be used to address the issue, is then demonstrated using the simulations. Lastly, we apply this method to actual data and discuss the practical importance of understanding the impact of spatial correlation structure and incommensurability.

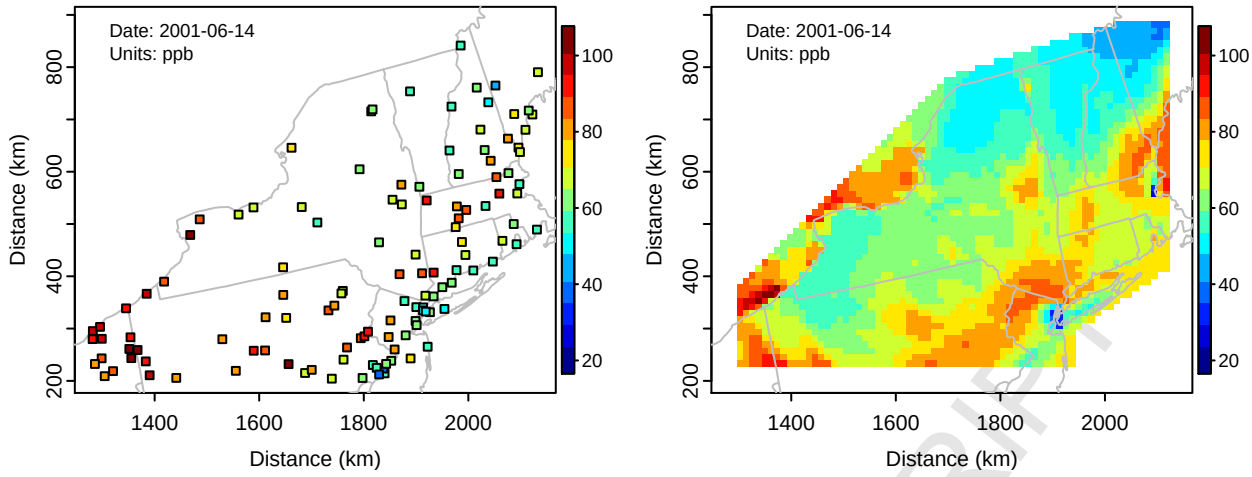
Key words: block kriging, air quality modeling, spatial interpolation, statistical simulation

1. Introduction

The performance of an air quality model is typically evaluated against actual measurements of the pollutant in question from monitoring networks. Such models treat the region as a large grid, giving output for each cell, so that the analyst must determine how to make the comparison. In the most common situation, the observed value from each monitor is matched with the value for the grid cell in which the monitor is located. The resulting paired data are used to evaluate the performance of the model, both visually, using such graphical displays as scatterplots and spatial plots, and numerically, using various performance metrics such as bias and root mean squared error. Examples of comprehensive air quality model evaluations which utilize this approach include Eder et al. (2006), Eder and Yu (2006), and Appel et al. (2007).

For example, consider observed and modeled maximum 8-hour ozone values in the northeastern United States on June 14, 2001. Fig. 1(a) shows the observations recorded at 124 air monitoring stations in the region in parts per billion (ppb). Output from a Community Multiscale Air Quality model (Byun and Schere,

*Corresponding author. Tel.: 919 541 7655; fax: 919 541 1379
Email addresses: Swall.Jenise@epa.gov (Jenise L. Swall), Foley.Kristen@epa.gov (Kristen M. Foley)



(a) Observed (b) Modeled

Figure 1: Observed and modeled maximum 8-hour ozone (2001-06-14)

2006) simulation for the same day is pictured in Fig. 1(b). This model run utilizes grid cells with each side of length 12 km. (Henceforth, we refer to grid cells by their side length, e.g. "12 km grid cell", "36 km grid cell"). As with many such models, the value for a grid cell represents the volume average for the layer of the atmosphere closest to the earth's surface over the extent of the grid cell. A scatterplot and summary statistics of the sort often used in traditional evaluation approaches are shown in Fig. 2. A spatial plot of the differences between the model-monitor pairs is given in Fig. 3.

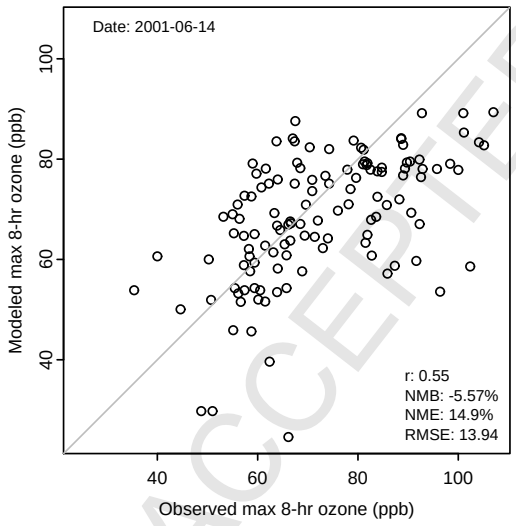


Figure 2: Modeled vs. observed maximum 8-hour ozone (2001-06-14)

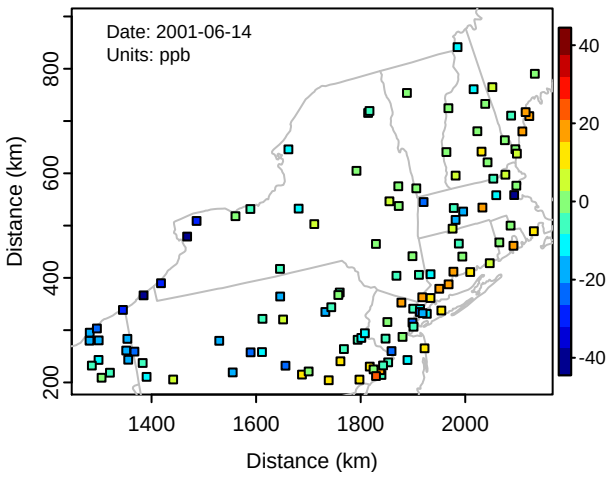


Figure 3: Differences (modeled-observed) in maximum 8-hour ozone (2001-06-14)

Figs. 1-3 are sufficient to form a general picture of model performance. An examination of Figs. 2 and

3 reveals that, for this particular day, the model underpredicts maximum 8-hour ozone at more monitoring sites than it overpredicts. Fig. 3 shows that while underprediction seems to be an issue along the Canadian border, we have a mix of overprediction and underprediction in other areas, particularly for coastal sites.

While these figures are helpful in understanding how the model output compares with the observations, none of them is sufficient for a detailed assessment of model performance. For instance, model metrics (such as those shown in Fig. 2) can only be calculated for grid cells in which we have monitors, and this means that the model metrics may overly reflect model performance in areas with large numbers of monitors. These might most often be urban areas or regions which have been pinpointed for further study due to a perceived greater likelihood of problems. None of these plots allows assessment of the model's performance at unmonitored locations. Also, the effect of measurement error or other sources of fine-scale variability cannot be adequately considered. Lastly, to better interpret Fig. 2, it would be helpful to understand to what extent the differences between model simulated values and observed values may be due to the inherent differences between point measurements and volume averages.

The problem of comparing grid averages and point measurements is usually termed "incommensurability" in the atmospheric science literature. Statisticians refer to this same issue as one of "change of support"; the issue and the underlying mathematics are discussed more thoroughly by Gelfand et al. (2001), while Gotway and Young (2002) give an extensive review of statistical methods for combining data with different spatial supports proposed for a variety of scientific applications. Although in the context of air quality modeling most evaluations have not considered this issue in detail, some authors have considered the problem and developed statistical methods for addressing it. The papers by Fuentes et al. (2003), Fuentes and Raftery (2005), Swall and Davis (2006), and Davis and Swall (2006) present sophisticated statistical models for various applications, all of which address the incommensurability issue and are able to estimate pollutant levels for grid cells in which no observations lie. Even so, since there are no consistently available, regionally comprehensive sources of observational data at heights beyond that of the typical monitoring station, each of these techniques treat the model output as areal, rather than volume, averages, and work herein follows suit.

In this paper, we discuss these issues from an applied statistical perspective. We introduce simulated datasets to explore the impact of various spatial correlation structures on common model evaluation tools. Using these simulated cases, we illustrate the benefit of the "block kriging" technique, which uses the observations and the spatial correlation among them to estimate the levels of a pollutant at all of the grid cells, whether or not they contain monitors. This technique has the advantage of being relatively easy to implement in statistical software packages and of requiring relatively few assumptions, compared with some of the more complex approaches developed by the above authors. We compare and contrast this technique with traditional, point-based kriging techniques. Lastly, we apply these ideas to selected real-life cases and

demonstrate their utility as part of a focused model assessment strategy.

2. Simulations

In this section, we make use of two sets of simulated spatial fields to demonstrate the potential impact of incommensurability on analyses. We also use these simulated examples in Section 3 to demonstrate the block kriging technique. One set contains cases in which, though the spatial correlation structures are different, the distance over which the spatial correlation extends is quite limited. The other set is made up of the same types of correlation structures, but with further-reaching spatial correlation. The correlation structures selected for this simulation study include three of the more commonly used models for spatial correlation: Gaussian, exponential, and spherical correlation models. Detailed descriptions of each of these models can be found in many standard geostatistical texts (e.g. Isaaks and Srivastava, 1989, pp. 373-375).

The simulations are performed using the RandomFields package (Schlather, 2001) within the R statistical computing environment (R Development Core Team, 2007). Each simulation is performed using the same example "region", a square area measuring 288 units on a side. The spatial field is generated over a dense lattice of 46 656 points. Although this discussion does not require the correspondence of our simulated domain with a particular modeling application, we envision the following simulations as taking place over a square domain with side length of 288 km, with 576 12 km grid cells.

2.1. Description

We focus first on cases in which the spatial correlation extends over long distances; we refer to these as simulations with "long-range" correlation. The black curves in Fig. 4 represent the correlograms (correlation vs. distance) for each of the three correlation structures under consideration. The correlation between points decreases as the distance between them increases, though the various correlation structures have a profound impact on the shape and rate of this decay. In contrast, the blue curves are cases of the same three structures, but in these the extent of spatial correlation is much shorter. We refer to this group as the short-range examples.

The range is defined as the distance at which the correlation between two points reaches zero, and Fig. 4 clearly shows a range of 360 km for the long-range spherical case and 90 km for the short-range. The Gaussian and exponential correlograms asymptotically approach zero, so that even at large distances some very small correlation is present. In such cases, the effective range is defined as the distance at which the correlation between points becomes negligible (typically < 0.05). For our long-range examples, the effective range is 360 km. So, given that the maximum distance (corner-to-corner) across the region is about 407 km, most of the points within the region are substantially correlated under the long-range correlation models. However, since the effective range for the short-range cases is much lower, 90 km, values of the spatial field in one portion of the region may be effectively uncorrelated with values in a different portion of the region.

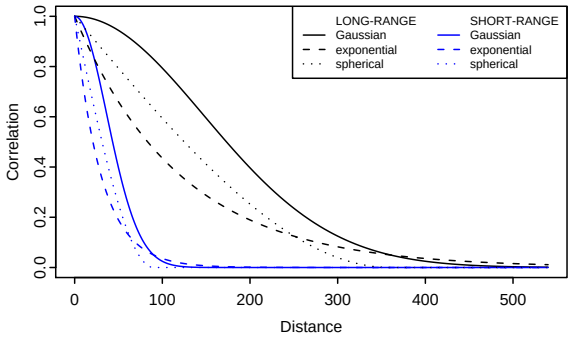
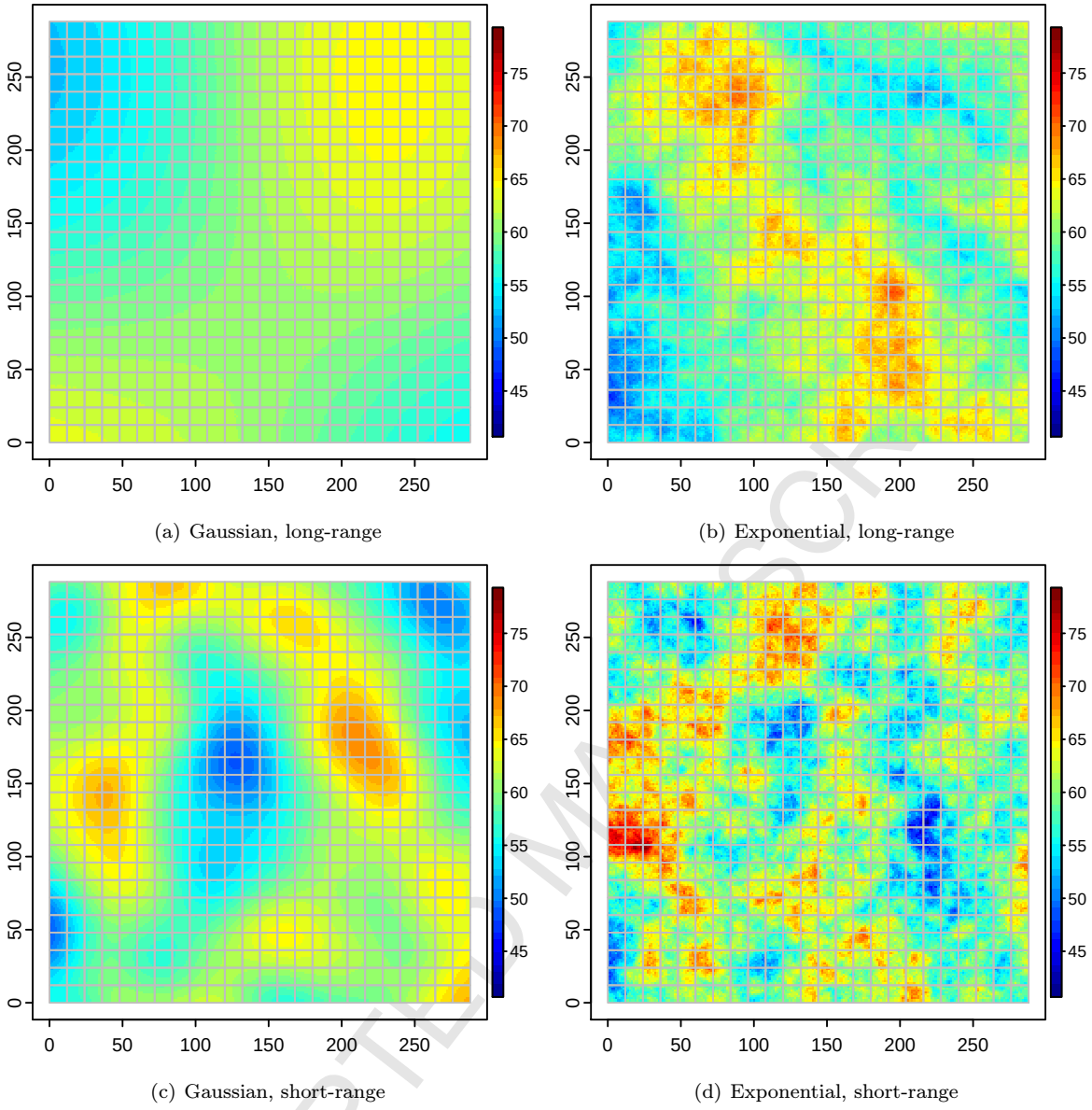


Figure 4: Correlograms for long-range and short-range simulations

We consider a few of the implications of these correlation structures by examining the sorts of spatial fields they produce. Fig. 5 shows example simulations for the Gaussian and exponential cases, based on the correlograms in Fig. 4. In each of these simulations, we hold constant the overall mean for the field and the partial sill (which gives the variance at any single location). The reader should note that these constitute only a sample of randomly generated simulations using the specified correlation structures, partial sill, and mean; additional simulations will yield spatial fields which look different, but share similar features of smoothness, variability, etc.

As shown in Fig. 4, the Gaussian correlation structure is distinguished by its slow and smooth reduction in correlation as distance grows, with correlations particularly high for small distances. This is reflected in the smooth gradations between areas of relatively low and high values in Figs. 5(a) and 5(c). The exponential correlation structure has a much sharper decrease in correlation for short distances, followed by a more gradual descent for larger distances; again, this is reflected in the much more variable spatial fields depicted in Figs. 5(b) and 5(d). As we might expect based on the correlograms in Fig. 4, the spherical case falls somewhere between the previous two, and we choose to focus for the remainder of the paper on the exponential and Gaussian cases to provide a greater contrast.

Using the dense fields in Fig. 5, we can average the values within each grid cell to obtain a representation of what the model output would be, assuming that the model is an accurate representation of the spatial field. (These averages are examined in detail for several cases in Sec. 3.) We also randomly select 28 locations to represent monitoring sites within this simulated region; these are shown in Fig. 6. This number is chosen for similarity to our motivating example, for which the ratio of monitors to grid cells is roughly 0.05. Together with the simulated model output, the selection of this sample of simulated observations allows us to investigate the performance of some commonly used evaluation strategies.



(a) Gaussian, long-range

(b) Exponential, long-range

(c) Gaussian, short-range

(d) Exponential, short-range

Figure 5: Gaussian and exponential simulations for the long-range and short-range cases

2.2. Examination of paired model output with observations

As discussed in Section 1, model evaluation may begin with a scatterplot of modeled values vs. observed values. Ideally, the points should lie close to the one-to-one line, with no trends to indicate potential biases. Fig. 7 displays such scatterplots for each of the simulated cases, assuming that there is no measurement or other localized error associated with each observation. This means that the observed values are simply the values of the dense field at the monitoring sites, so that any variation around the one-to-one line in Fig. 7 is due solely to incommensurability.

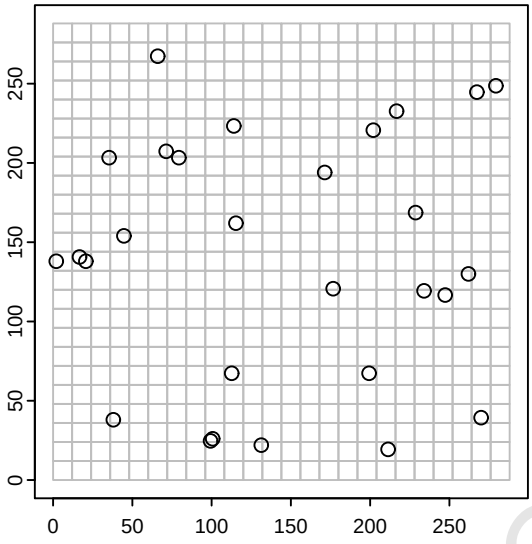


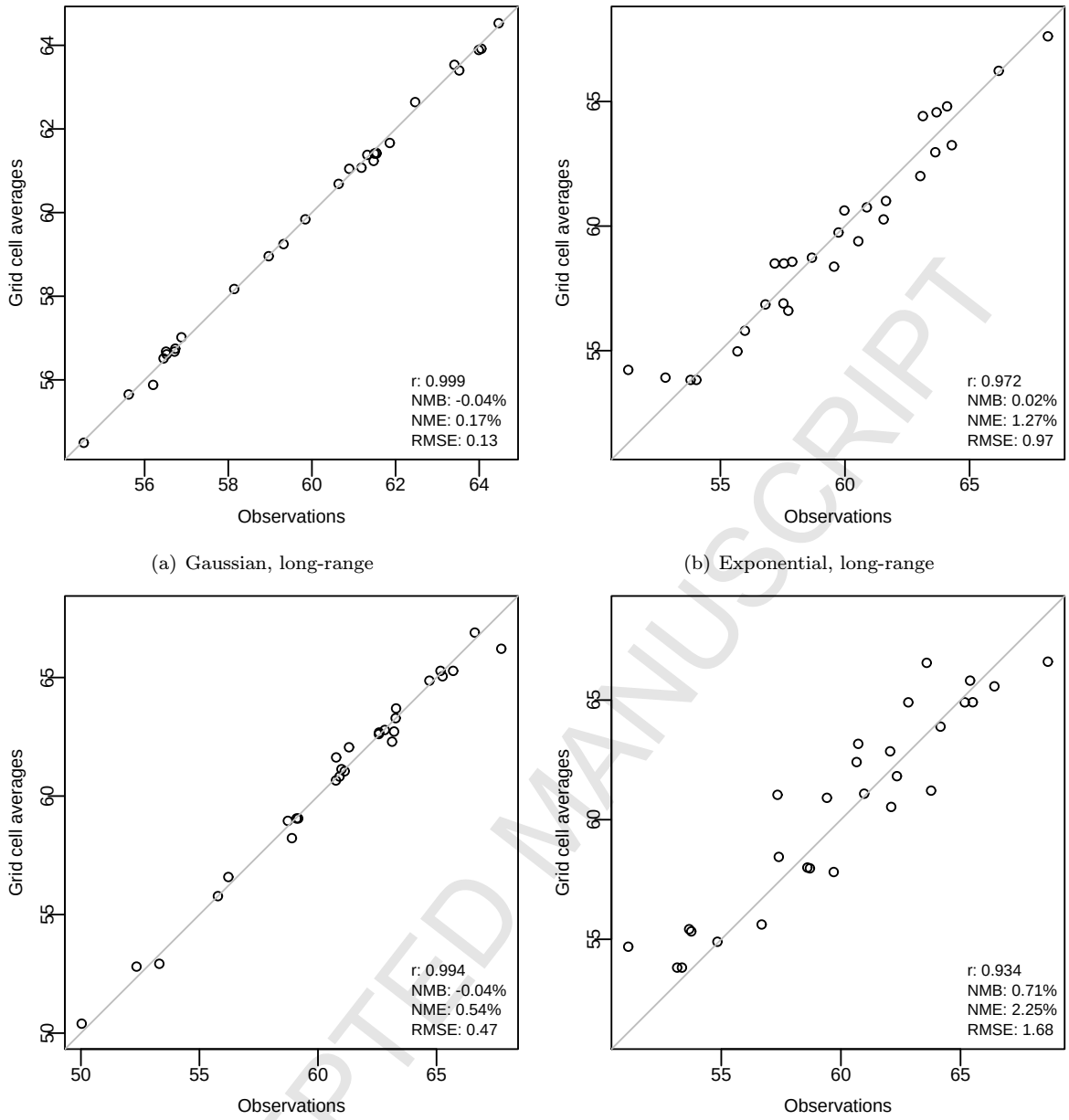
Figure 6: Simulated observation locations

In none of these scatterplots do all the points lie on the gray one-to-one line, though they are clustered around it. In the Gaussian case, in which the decrease in correlation is very slow for small separation distances, the correlation coefficient, r, is close to one in both the long-range and short-range cases. However, in the exponential case, the effect of incommensurability is more pronounced, especially for the short-range case, in which r is smallest.

To better understand the effects of incommensurability, consider the correlation in values taken at opposite corners, the most distant points in a 12 km grid cell. Table 1 shows that for the exponential correlation structure, this value is about 0.57 for the short-range case, while for the long-range case, it is only about 0.87. This is in contrast with the smooth descent of the Gaussian correlation function for short distances, in which the correlation between opposite corners of a grid cell is about 0.90 in the short-range case and 0.99 in the long-range case. It is of interest to note that because of differences in the shapes of the Gaussian and exponential correlograms at very small distances (Fig. 4), the short-range Gaussian correlogram yields higher correlation between opposite corners of a 12 km grid cell than the long-range exponential correlogram, in which the correlation dies off more quickly after the initial descent.

Although our simulated examples utilize 12 km grid cells, lower resolution model output is not uncommon. For instance, 36 km grid cells may be used for air quality model runs for a large region, such as the contiguous 48 states of the United States (e.g. Eder and Yu, 2006). In the case of larger grid cells, the greatly reduced correlations in the last column of Table 1 show that incommensurability plays a much larger role and can be expected to have a large impact in cases with effective ranges similar to those used in our short-range simulations.

A variety of metrics, in addition to the correlation coefficient, are widely used in model evaluation (e.g.



(a) Gaussian, long-range

(b) Exponential, long-range

(c) Gaussian, short-range

(d) Exponential, short-range

Figure 7: Scatterplots of simulated model output vs. observations (no measurement error)

Eder and Yu, 2006; Eder et al., 2006). A few such common summary statistics, such as normalized mean bias (NMB), normalized mean error (NME), and root mean square error (RMSE), are included in the lower right corners of the scatterplots in Fig. 7. These also indicate the presence of some error, even though these cases were simulated assuming a "perfect" model and observations. This demonstrates that traditional assessment tools may yield results which are hard to interpret, particularly when the correlation structure has a rapid

Table 1: Correlation between values at diagonally opposite grid cell corners

Extent	Structure	12 km^2 cell	36 km^2 cell
Long-range	Gaussian	0.99	0.94
Long-range	exponential	0.87	0.65
Short-range	Gaussian	0.90	0.38
Short-range	exponential	0.57	0.18

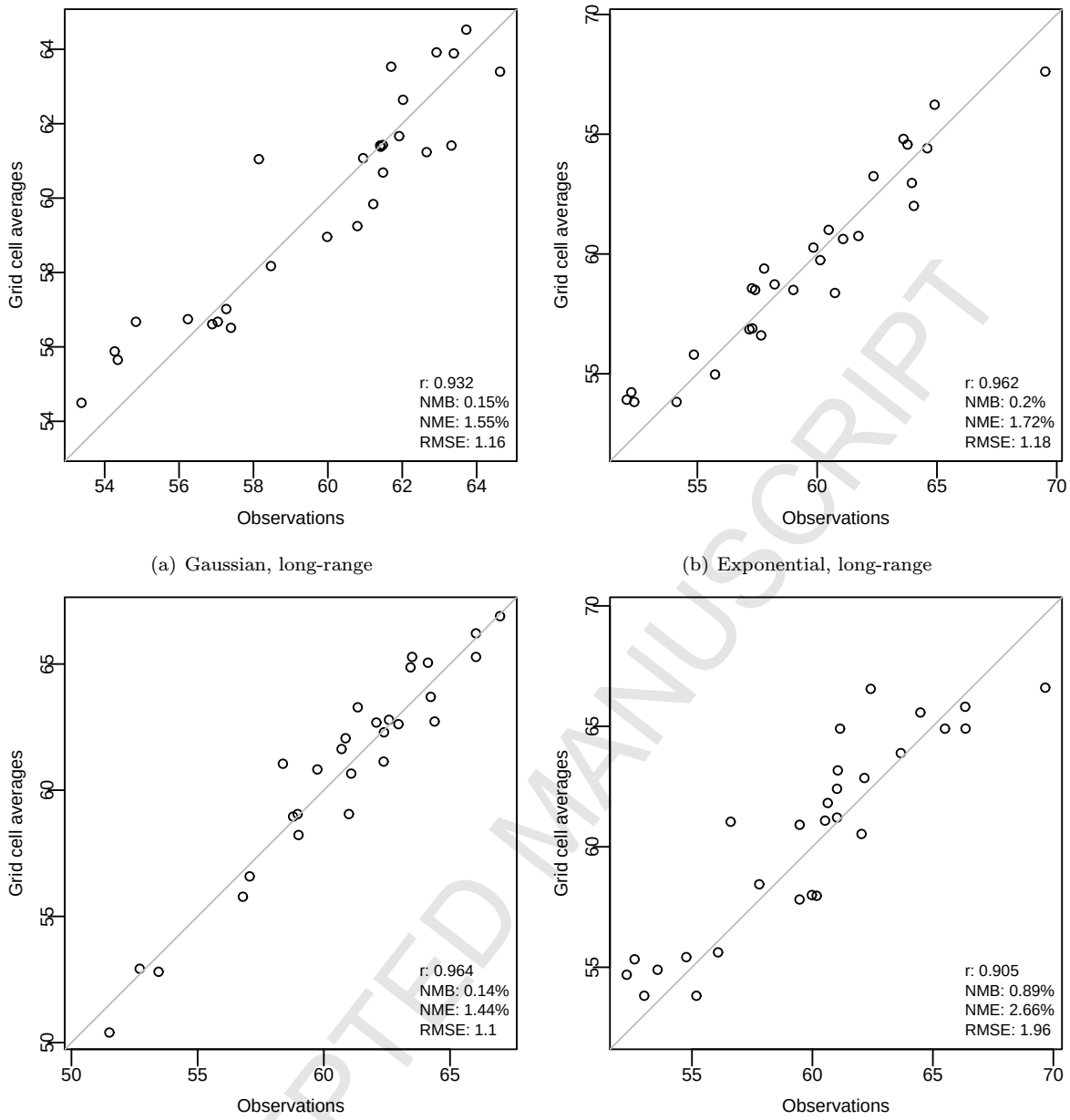
decrease for small separation distances (as opposed to the very gradual descent of the Gaussian case) and/or when the effective range is short.

An additional complication in real-life applications is the presence of measurement or other fine-scale error, however small, which adds to the variability around the one-to-one line. To see this effect, we add normally distributed error with standard deviation one to the observations. Fig. 8 shows the resulting scatterplots. We see that the addition of even this small amount of variability, together with incommensurability, makes it more difficult for the viewer to discern whether the model's performance is acceptable for an envisioned application.

3. Kriging techniques for model evaluation

In Section 2.2, we performed comparisons of observations with model output using scatterplots and evaluation metrics for our simulated data. Such analyses have the advantage of being straightforward to conduct, while providing some helpful quantitative measures of performance. However, using such methods, we can only evaluate model performance for grid cells in which monitoring sites are located, which means that, in most cases, the majority of the grid cells cannot be evaluated. In addition, we have no way to determine whether output at a particular model grid cell is "close enough" to the observation that we would attribute the difference to the incommensurability issue, rather than a potential problem with the model. To address these points, we make use of kriging techniques to estimate the level of the pollutant in question for each grid cell in the region of interest, using the simulated observations. The resulting estimates can then be compared with the simulated model output.

While other methods for spatial interpolation are available, kriging techniques are able to make use of the extent and nature of spatial correlation present in the region to improve the accuracy of the resulting estimates. This spatial correlation structure and associated parameters are often estimated based on the observational data, and are usually assumed to be constant over the focus region. In addition, the kriging methodology provides the standard errors associated with its estimates. These error estimates aid in the determination of whether a difference between the kriging surface and the model output is likely due to



(a) Gaussian, long-range

(b) Exponential, long-range

(c) Gaussian, short-range

(d) Exponential, short-range

Figure 8: Scatterplots of simulated model output vs. observations (with measurement error)

error attributable to the statistical technique or to a possible discrepancy between the modeled and observed levels. Further details about the theory and application of kriging techniques are widely available in the literature; see, for instance, Cressie (1993) and Isaaks and Srivastava (1989).

Admittedly, more sophisticated statistical approaches fulfill these same objectives, and offer additional flexibility for better variability estimates, incorporation of other relevant information, and other model

improvements. Several such methods are mentioned in Sec. 1. The benefits of the block kriging approach discussed here are mainly found in the ease with which it can be implemented in available software packages, the size of the regions for which it can be applied without a substantial investment in computational time or resources, and the relatively short spin-up time required to learn the methodology. One strategy would be to use traditional techniques (scatterplots, expert opinion, etc.) to identify potential modeling problems, and then to apply block kriging in these cases. As concerns are better identified, appropriate statistical procedures can be used to assess model performance more precisely.

3.1. Block kriging

Our goal is to estimate the average level of the spatial field for an entire grid cell, and not just the level at a particular point in the grid cell. The strategy of simply kriging to the cell centers does not account for this incommensurability. Instead, we use block kriging, as described by Goovaerts (1997, pg. 152), Isaaks and Srivastava (1989, Chp. 13) and others, to adjust for the incommensurability between points and cell averages. The technique involves kriging to a grid of points within each cell (each "block"), and using the sample mean of the kriging estimates at these points to estimate the cell average. Representing each kriging point estimate on this grid as X_i and assuming a grid of size n, the block kriging estimate is the sample mean of the estimates at each grid point as follows:

$$\text{block kriging estimate} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \tag{1}$$

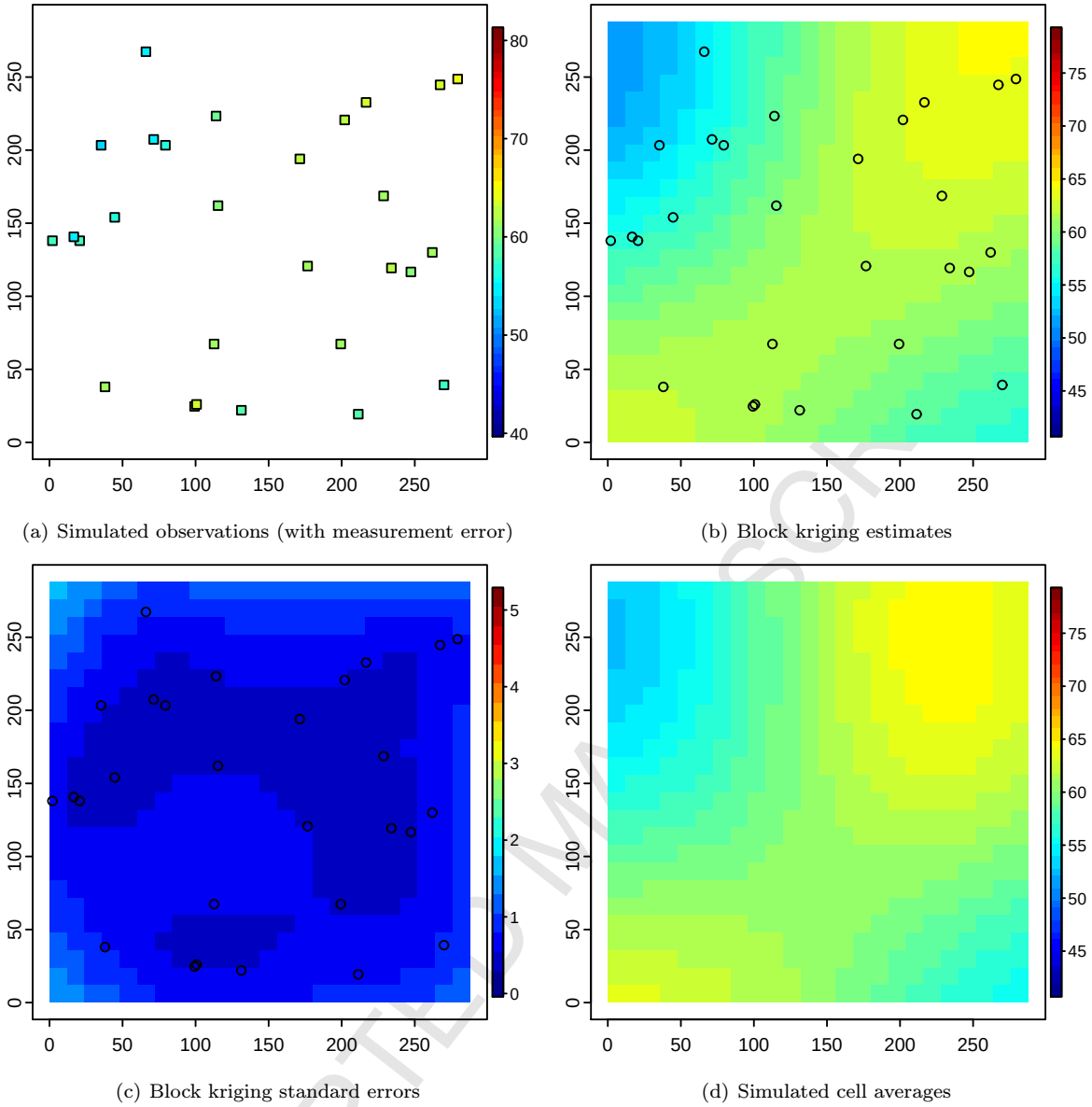
The standard error (SE) of the estimate for the cell average is then the square root of the variance of the sample mean in Eq. 1. Since the kriging estimate for each of the grid points is correlated with the others, this calculation must take the covariances between kriging estimates at each of the n grid points into account.

$$\text{SE of estimate} = \sqrt{\text{Var}(\bar{X})} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n^2}\text{Cov}(X_i, X_j)} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\text{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i\neq j}}^{n}\text{Cov}(X_i, X_j)} \tag{2}$$

The block kriging estimates and standard errors can be obtained using statistical software packages; we used the gstat package (Pebesma, 2004) in the R statistical computing environment.

We demonstrate this technique using the simulated observations with error, as described at the end of Section 2.2. Since the correlation structure used to generate the simulated data is known, we are saved the additional step of having to estimate the correlation structure from the data. This is an advantage which we would be unlikely to have in an actual application, as seen in our presentation of real-life examples in Sec. 4.

Fig. 9 gives a complete picture of the kriging process for the long-range Gaussian case. Block kriging takes the simulated observations in Fig. 9(a) as input, returning the estimates in Fig. 9(b) and the associated standard errors in Fig. 9(c). In the latter two plots, the observation points are denoted by black circles for quick reference. Of our simulated cases, only in this long-range Gaussian example does kriging provide a



(a) Simulated observations (with measurement error)

(b) Block kriging estimates

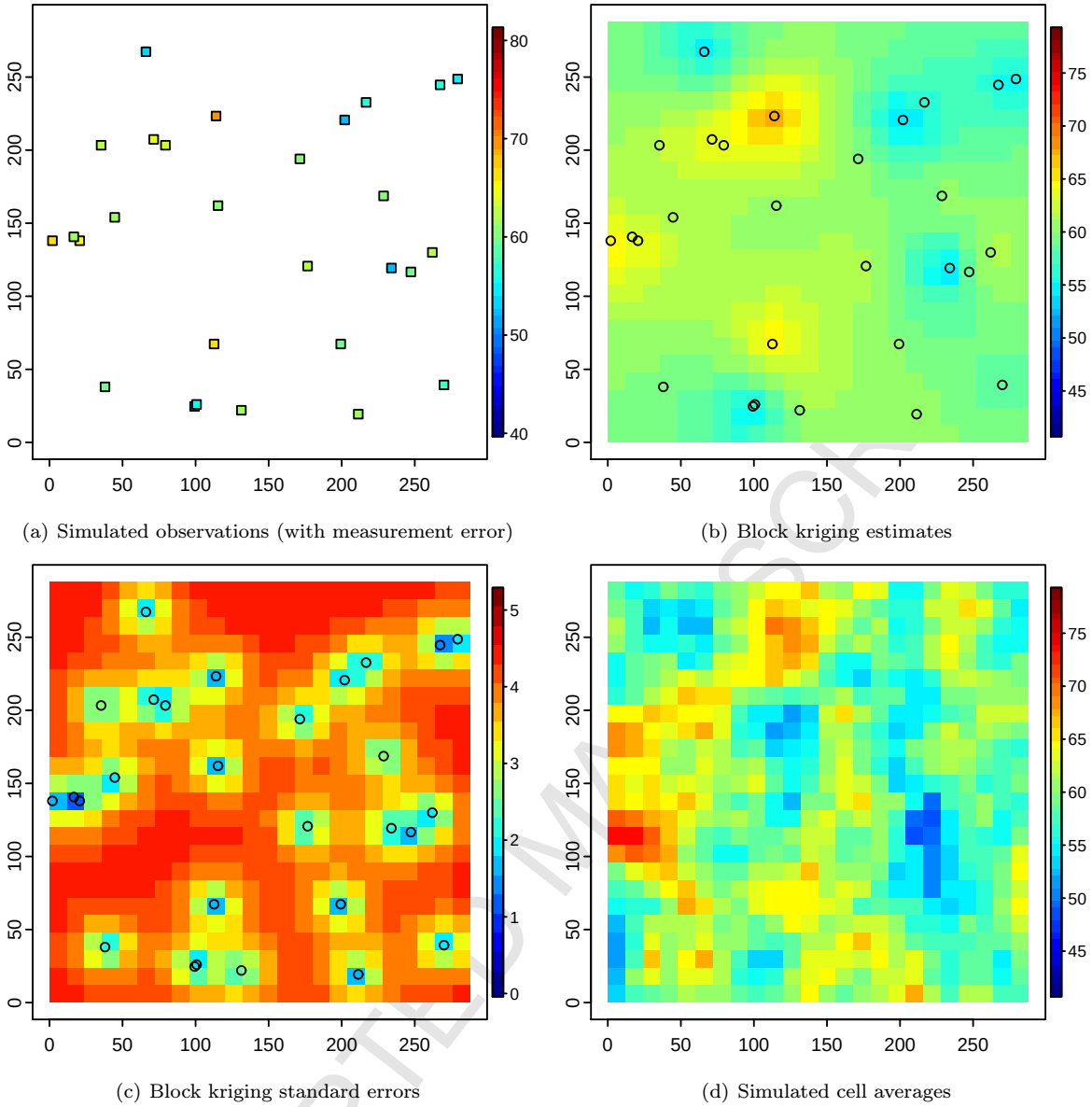
(c) Block kriging standard errors

(d) Simulated cell averages

Figure 9: Kriging process (Gaussian, long-range)

visually representative picture of the actual cell averages shown in Fig. 9(d). The underlying spatial field in this case (Fig. 5(a)) has much less variability than in the other three simulations, with very gradual changes. This is due to the smooth change in correlation as distance increases and the large effective range, as seen in the long-range Gaussian correlogram in Fig. 4. These factors contribute toward making the Gaussian long-range case easiest to estimate, as reflected by the low standard errors in Fig. 9(c), which range from 0.4-1.6 ppb.

As expected, the hardest estimation situation of the four considered is that pictured in Fig. 10. The



(a) Simulated observations (with measurement error)

(b) Block kriging estimates

(c) Block kriging standard errors

(d) Simulated cell averages

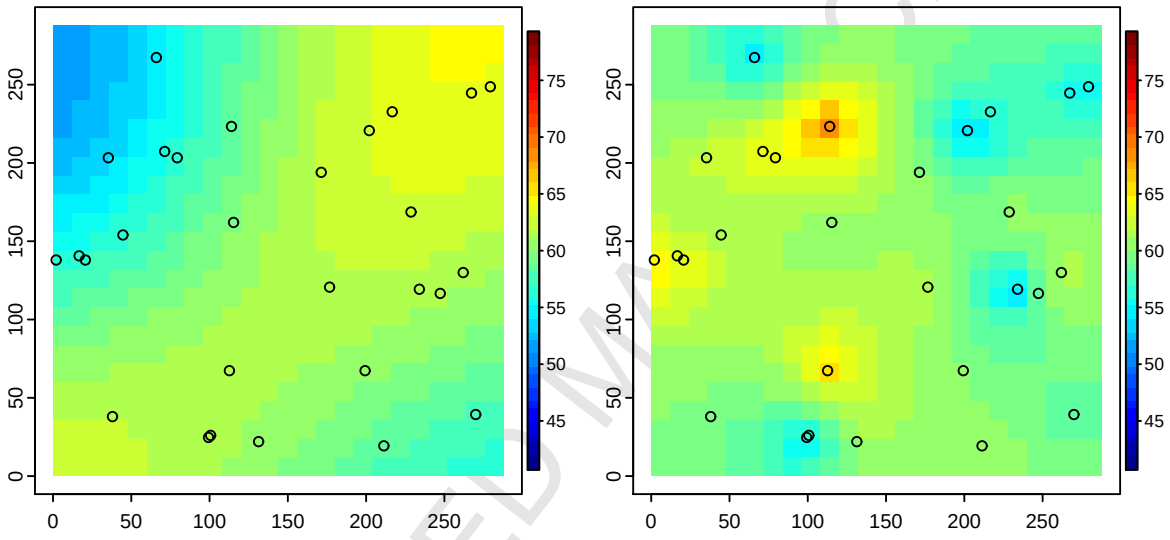
Figure 10: Kriging process (exponential, short-range)

short-range exponential correlation structure allows greater variability and more extreme values (as seen in Fig. 5(d) and the simulated cell averages in Fig. 10(d)). The kriging estimates in Fig. 10(b) are oversmooth, with estimates tending toward the overall mean in areas in which the distances to the nearest observations are nearing the effective range. To reflect the lack of information in these cases, the standard errors increase well beyond the levels of error seen in the more well-behaved long-range Gaussian case (Fig. 9), ranging from 1.2-4.5 ppb. In cases such as this one, in which the spatial correlation dies off quickly, more observations are needed to capture the field adequately.

The long-range exponential and short-range Gaussian cases are less extreme than the cases in Figs. 9 and 10. They are not pictured here due to space limitations. In both cases, the kriging estimates capture the general pattern, but are not able to correctly place the locations or intensities of many of the extreme values.

3.2. Comparison of block kriging vs. point kriging

We have proposed block kriging as a method that both allows us to assess the performance of the model at grid cells in which no monitors lie and accounts for the incommensurability between point measurements and grid cell averages. This is in contrast with the simpler technique of simply kriging to the center of each grid cell, and using the estimates at the cell center points as estimates for the cell averages. Again, we focus on the two most extreme cases, the long-range Gaussian and short-range exponential cases.



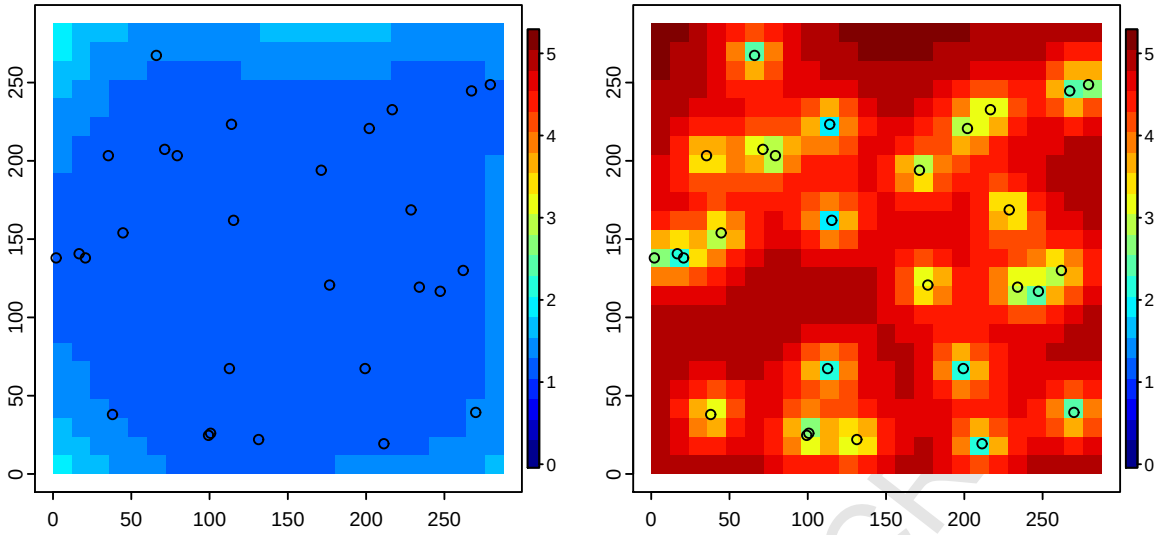
(a) Gaussian, long-range

(b) exponential, short-range

Figure 11: Estimates using kriging to cell centers

For all our simulated examples, the correlation between pollutant levels at locations within the same grid cell is still reasonably high (at least 0.57, see Table 1). For most grid cells, then, the estimate for the the grid cell average will not differ much from the estimate at the center point. This can be seen by a visual comparison of Fig. 11 with Figs. 9(b) and 10(b). However, the standard error associated with the grid cell average may be much lower than that for any single point within the grid cell, particularly in cases in which the correlation dies off rapidly in the short distance across the grid cell. This is due to the smaller contributions (substantially less than one) from the covariance terms in Eq. 1.

Fig. 12 shows the standard errors we obtain using kriging to the cell centers, for comparison with the block kriging errors in Figs. 9(c) and 10(c). In both cases, we see that the standard errors associated with



(a) Gaussian, long-range

(b) exponential, short-range

Figure 12: Standard errors using kriging to cell centers

kriging to the cell centers are larger than those for block kriging. In the long-range Gaussian case, these standard errors are 0.3-0.7 units higher than those for the block kriging case. In the short-range exponential simulation, the standard errors associated with kriging to the cell centers are 0.3-1.0 units higher.

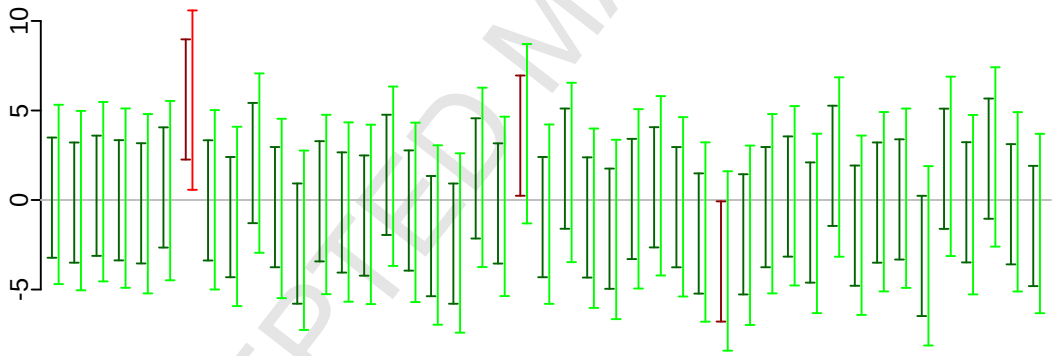


Figure 13: 95% confidence intervals associated with block kriging and kriging to the cell center. The left interval in each pair (colored dark green or dark red) corresponds to block kriging; the right interval is formed by kriging to the cell center.

The kriging estimates and standard errors can be used to build 95% confidence intervals for each grid cell of the form

$$\text{kriging estimate} \pm (1.96 \times \text{SE of estimate}) \tag{3}$$

For approximately 95% of randomly-generated simulated fields with the specified covariance structure and the same placement of monitors and grid cells, the confidence interval for a given grid cell should enclose the

simulated cell average. Because there is less variability associated with statistical estimates of an average, the standard error in Eq. 2 leads to confidence intervals which are better calibrated to the desired confidence level. As an illustration, we examine the calibration of the confidence intervals by generating 500 randomly simulated fields using the short-range exponential correlation structure described before.

For a particular grid cell, Fig. 13 displays pairs of confidence intervals obtained for a randomly chosen subset of 45 of these 500 simulations. Each pair of confidence intervals has been re-centered around zero; this is necessary for comparison purposes since the cell average is different in the various simulations. Confidence intervals which enclose the target value are shown in green, and those which do not are shown in red. The left interval in each pair (colored dark green or dark red) corresponds to block kriging, while the interval on the right corresponds to kriging to the cell center. In 95.2% of our 500 simulations, the block kriging confidence interval enclosed the target mean, but for the intervals yielded by kriging to cell centers, 98.8% did. For other grid cells, these numbers are slightly different, but in almost every case the success rate of the block kriging intervals was closer to 95%. The smaller standard errors, and thus shorter confidence intervals, associated with block kriging are a better reflection of the statistical variability associated with estimates of grid cell averages. The implications for model evaluation are discussed in the following section.

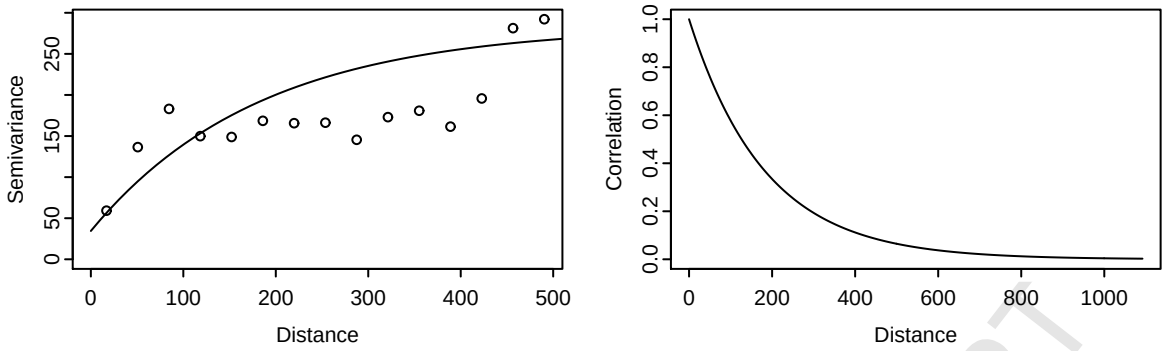
4. Comparison of kriging techniques using actual data

Having demonstrated the utility of kriging and the impact of incommensurability on error estimates using simulation case studies, we explore the applications of these techniques to actual air quality applications. We highlight daily maximum 8-hour ozone concentrations in the northeastern U.S. on two days during the summer of 2001, with one day exhibiting a longer-range spatial correlation and the other with a more moderate effective range. In both cases, the monitoring data is derived from the U.S. Environmental Protection Agency's Air Quality System (AQS) sites and the model output from the Community Multiscale Air Quality (CMAQ) system (Byun and Schere, 2006). This simulation used CMAQ version 4.5, with 12 km grid cells. Meteorological input came from MM5 (Grell et al., 1995), and emissions information came from the 2001 National Emissions Inventory.

4.1. Long-range example: June 14, 2001

We return to the example presented in the introduction from June 14, 2001. To save space, we refer the reader to Fig. 1-3 to visualize the monitoring data, model output, and paired differences for this example.

Based on the observed ozone concentrations in Fig. 1(a), we use the kriging techniques discussed in Section 3 to obtain ozone estimates for each of the model grid cells in Fig. 1(b). Unlike our simulated cases, our correlation structure is unknown and has to be estimated based on the observed data. We do this through variogram modeling and restricted maximum likelihood estimation (REML), as described by authors such



(a) Empirical (points) and estimated (line) variograms

(b) Estimated correlogram

Figure 14: Variograms and correlogram (2001-06-14)

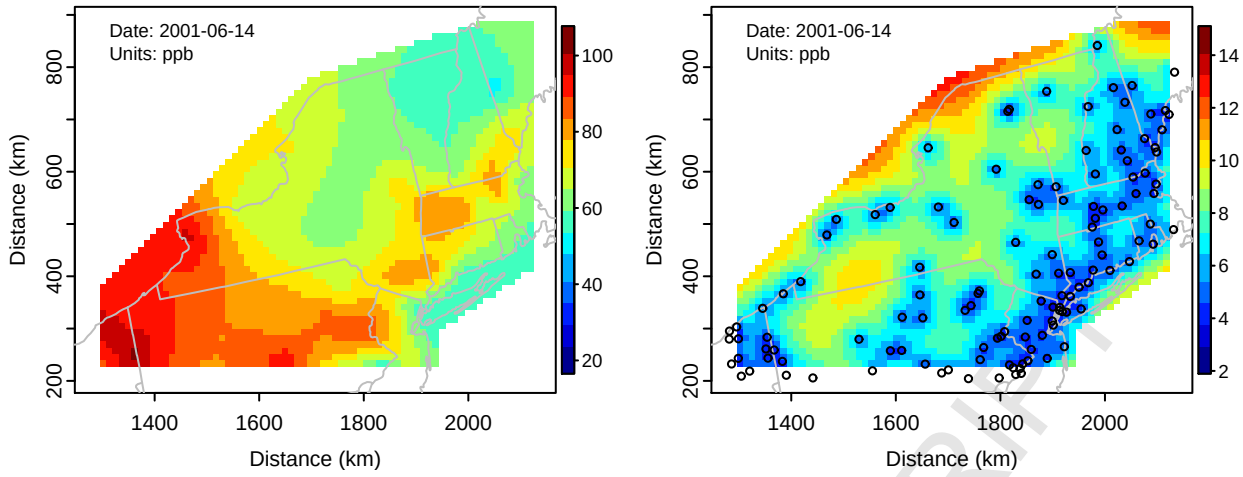
as Cressie (1993, Chp. 2.6), who provides a detailed treatment of the subject. The estimated variogram (solid line), superimposed on the empirical variogram (points), is displayed in Fig. 14(a). Fig. 14(b) shows the resulting estimated correlogram, which has an exponential correlation structure and an effective range of approximately 547 km, roughly 1.5 times the effective range used in on our long-range simulations.

The REML estimates also include a partial sill of about 249 ppb^2 and fine-scale error with approximate variance 34.5 ppb^2. The large estimated partial sill is reflected in the variability of the observations in Fig. 1(a), which might initially give the impression that the correlation structure is a short-range one. The estimation of covariance structures is notoriously difficult, and kriging methodologies do not account for the error inherent in the process. In cases in which the data or other available information is insufficient to make a good estimate, or in cases where the standard error estimates are of particular importance, the use of more sophisticated Bayesian kriging techniques is advisable (e.g. Handcock and Stein, 1993; Swall and Davis, 2006).

Fig. 15 displays the block kriging estimates and associated standard errors. Comparison of Figs. 1(b) and 15(a) shows that the kriging estimates are spatially smoother than the CMAQ-simulated surface, which is not surprising, giving the weighted-averaging that is the basis of the technique. The standard errors are given in Fig. 15(b), ranging from 3.1-12.8 ppb. As previously noted, the estimated standard errors are much lower in the vicinity of monitors, and this is especially notable along the eastern seaboard. Regions near the Canadian border and an interior southwestern portion of New York and Pennsylvania have fewer monitors, and thus much higher uncertainties.

Based on our earlier discussion, we would expect that failing to adjust for incommensurability would have only minor impacts on the kriging surface, and Fig. 16(a) shows that the surface obtained by kriging to the grid cell centers is similar to that estimated using block kriging (Fig. 15(a)). However, the standard error estimates associated with kriging to the cell centers are on average 1.4 times larger, as seen in Fig. 16(b).

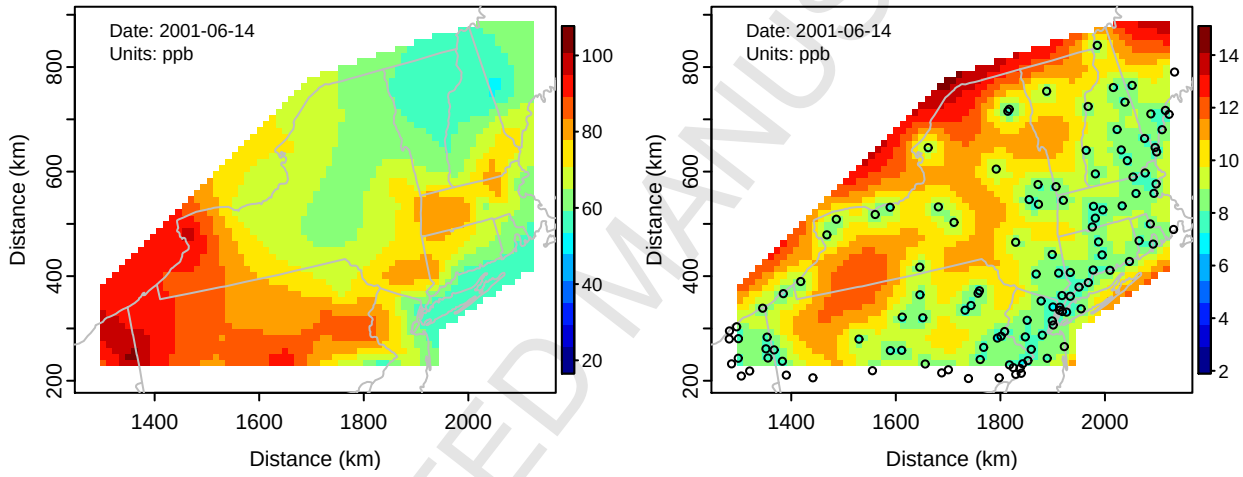
As in Sec. 3.2, we can use the kriging estimates and standard errors in Fig. 15 to build a 95% confidence



(a) Estimates for grid cell averages

(b) Estimated standard error

Figure 15: Block kriging estimates (2001-06-14)



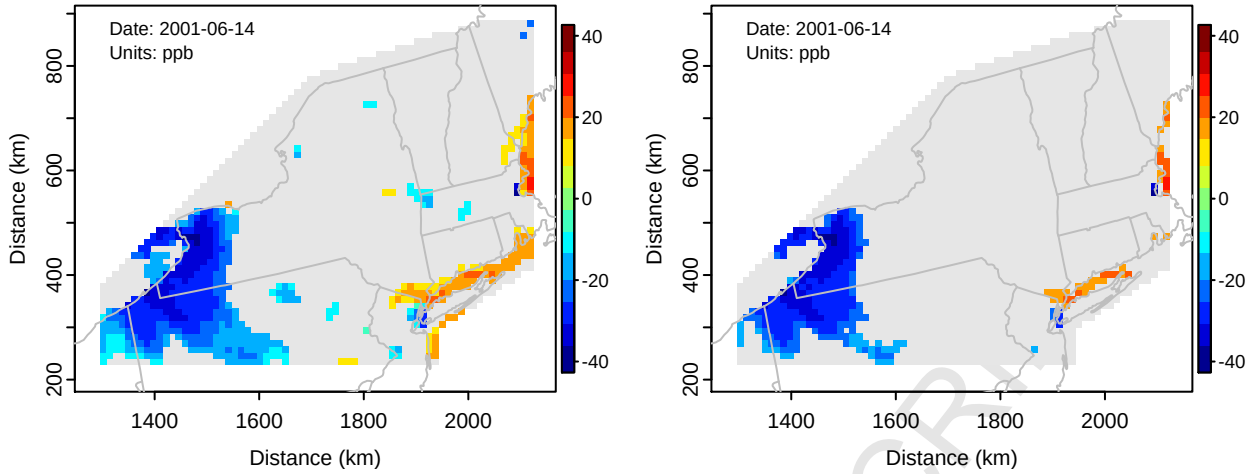
(a) Kriging at grid cell centers

(b) Estimated standard error

Figure 16: Kriging to cell centers estimates (2001-06-14)

interval for each grid cell based on Eq. 3. We then inspect the model output to determine which grid cells, if any, fall outside the bounds given by the confidence interval. The difference between the model-simulated value and the block kriging estimate based on the observations is then said to be statistically significant, and these grid cells are deemed candidates for further, more detailed, inspection. Note that we cannot say that the model is "wrong" for these grid cells, since there are additional factors that may explain the differences. For instance, CMAQ is able to make use of meteorological and emissions information, which is not always captured by the sparse monitoring networks. Also, since these are 95% confidence intervals, even if the model were performing perfectly and if there were no fine-scale or measurement error, we would expect about 5%

of the grid cells to have model-simulated values outside the confidence bounds.



(a) Block kriging

(b) Kriging at grid cell centers

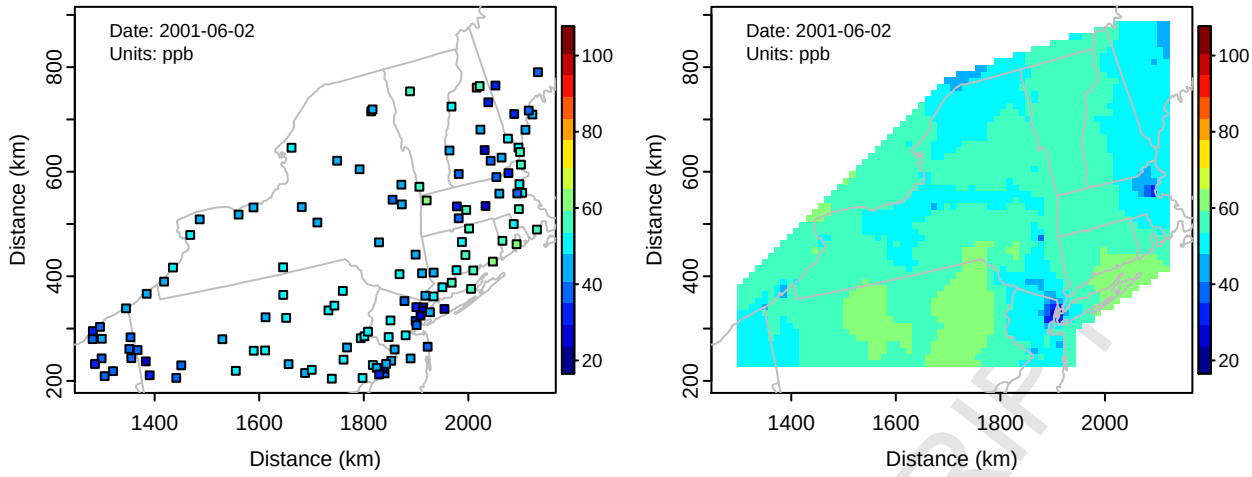
Figure 17: Locations of grid cells of interest for 2001-06-14

The colored grid cells in Fig. 17(a) denote the grid cells whose simulated value falls outside the 95% confidence intervals obtained by block kriging. The color gives the difference (in ppb) between the modeled value and the block kriging estimate for these cells, while any differences which are not significant are shaded gray. Fig. 17(a) shows the substantial underprediction of maximum 8-hour ozone in the southwestern corner of the region, including northwestern Pennsylvania and southwestern New York. In addition, we see some overprediction along the coast, from New Jersey to Maine.

If we were to neglect incommensurability, we would construct confidence intervals based on kriging to the cell centers, rather than block kriging. Since we can estimate the average concentration over a grid cell with more precision than the concentration at the center point of the grid cell, the confidence intervals associated with block kriging will be narrower. This means that larger differences between modeled values and kriging estimates will be required to attain significance, so that block kriging has more "skill" in finding areas in which CMAQ and the estimates based on the observations are in disagreement. Fig. 17(b) shows the significant differences found when kriging to the cell centers is used. While these differences are in general agreement as to the areas worthy of further investigation, the block kriging technique better identifies the extent of the problematic regions.

4.2. Short-range example: June 2, 2001

We now turn to an example in which the effective range of the spatial correlation is substantially shorter. Fig. 18 shows the daily maximum 8-hour ozone measurements and model output for June 2, 2001. We notice immediately that for most of the region, this was a low ozone day. This is particularly notable in the



(a) Observed (b) Modeled

Figure 18: Observed and modeled maximum 8-hour ozone (2001-06-02)

southwestern portion of the region and in a few scattered locations elsewhere. The modeling data reflect higher values than do the monitors throughout the majority of the region. The paired model-simulated vs. monitored values are shown in the scatterplot Fig. 19. This, along with a few typically used measures of model performance displayed at the bottom right, confirms our initial impression that the model is overpredicting ozone for most monitoring locations in the region.

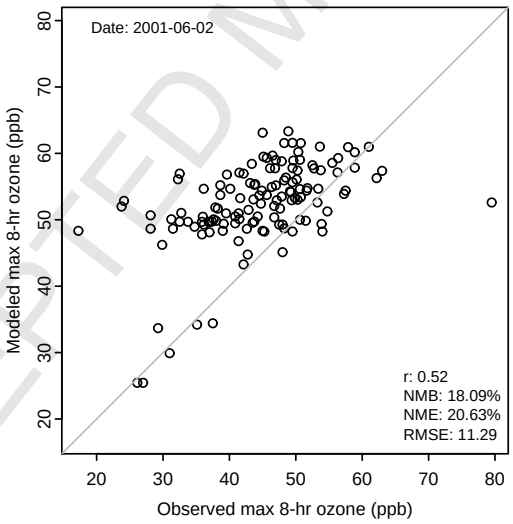
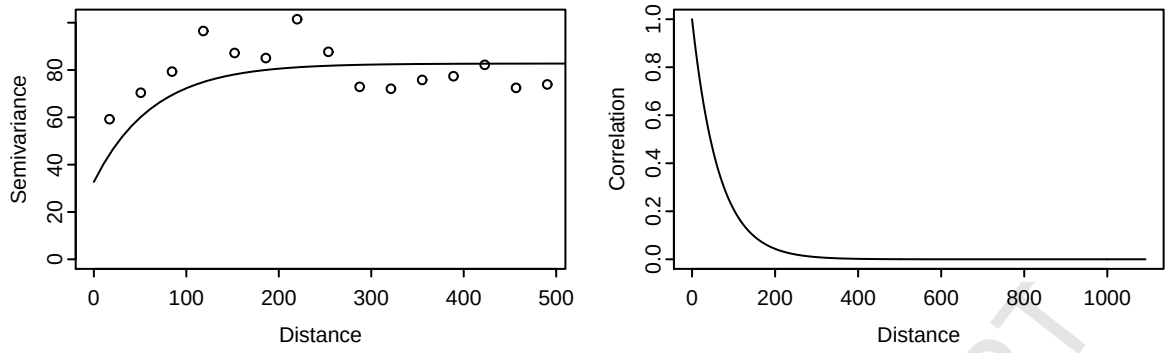


Figure 19: Modeled vs. observed maximum 8-hour ozone (2001-06-02)

Based on the observed ozone concentrations in Fig. 18(a), we use the block kriging techniques to obtain ozone estimates for each of the model grid cells in Fig. 18(b). Again, the correlation structure and associated parameters are estimated using REML based on the observed data, resulting in the variograms and correl-

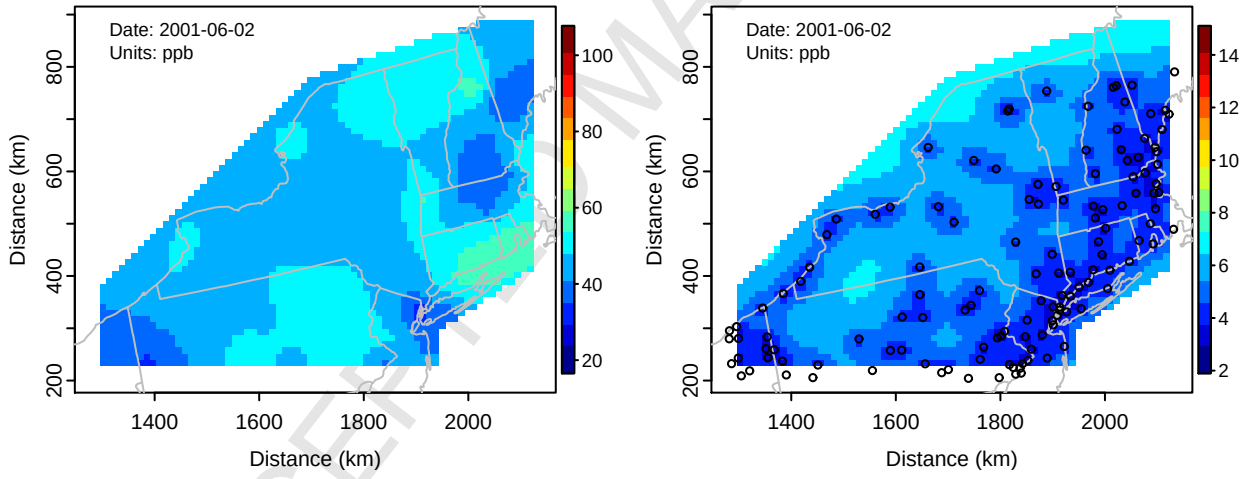


(a) Empirical (points) and estimated (line) variograms

(b) Estimated correlogram

Figure 20: Variograms and correlogram (2001-06-02)

ogram shown in Fig. 20. While this correlogram is exponential like that for the previous case (Fig. 14), it has an effective range of only about 191 km. The REML estimate for the partial sill is also much lower than that for the previous example, even with the shorter effective range of spatial correlation. This reflects the lower variability of the ozone concentrations observed on this day, compared to the previous example. Since the effective range is about 35% of the effective range of the previous case, we expect incommensurability to play a greater role.



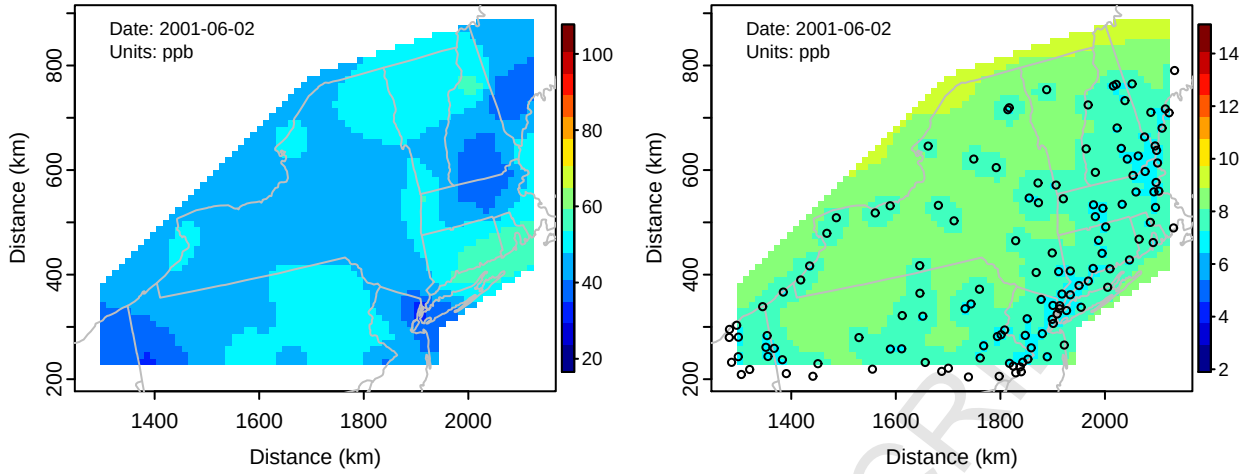
(a) Estimates for grid cell averages

(b) Estimated standard error

Figure 21: Block kriging estimates (2001-06-02)

Ozone estimates and standard errors obtained using block kriging are displayed in Fig. 21. As expected, the block kriging results are generally lower than the model simulated values. Again, the smoothing inherent in the kriging procedure yields estimates with less texture than we might believe to be realistic. The standard errors range from about 2.7-6.8 ppb, which is smaller than the range of standard errors for the

previous example. This difference is largely due to the smaller partial sill in this example.

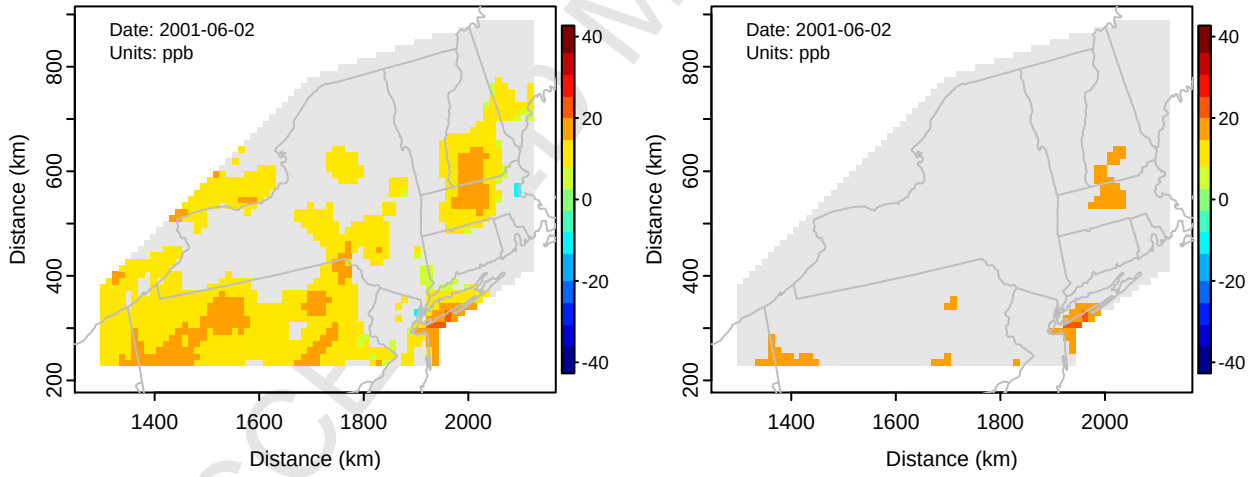


(a) Kriging at grid cell centers

(b) Estimated standard error

Figure 22: Kriging to cell centers estimates (2001-06-02)

Kriging surfaces and associated errors obtained by kriging to the cell centers are shown in Fig. 22. These estimates are once again nearly identical to those given by block kriging. The effect of our failure to account for incommensurability can be clearly seen in Fig. 22(b), which shows standard errors on the order of 6.6-9.1 ppb, on average about 1.6 times higher than those obtained using block kriging.



(a) Block kriging

(b) Kriging at grid cell centers

Figure 23: Locations of grid cells of interest for 2001-06-02

Fig. 23 highlights the grid cells whose model-simulated value falls outside the 95% confidence bounds given by the two types of kriging. As before, the color indicates the difference between the modeled value and kriging estimate, and gray indicates grid cells in which this difference was not statistically significant.

While both plots identify potential areas of overprediction in southwestern Pennsylvania, the vicinity of Long Island, and an area in central Massachusetts/New Hampshire, the block kriging methodology is better able to show the likely extent of these differences. Fig. 23(a) shows that grid cells in most of Pennsylvania, portions of western and central New York, the vicinity of Long Island, and large portions of Massachusetts and New Hampshire should all be investigated further to determine the cause of the apparent overprediction.

5. Discussion

Our simulations show that an understanding of both the type and extent of the spatial correlation structure can greatly inform the model evaluation process. We cannot assume that a measured value at a monitoring location within a cell is representative of the cell average, especially when the effective range is very short relative to the width of the modeled grid cells. This means that, even if the model is performing perfectly and there is no observational error, we cannot expect that in a scatterplot, points representing paired modeled and observed values will lie on a one-to-one line. Our comparison of Gaussian and exponential correlation structures with the same effective range shows that this concern looms larger for correlation structures in which there is a rapid decrease in correlation for small distances relative to grid cell size (like the exponential).

Kriging methods allow us to use the information contained in the monitoring data and information about the spatial correlation structure of the field to make estimates for unmonitored locations. This allows model assessment to proceed for modeled areas in which no monitors are sited. The discussion in Sections 3 and 4 shows that the incommensurability issue impacts our choice of kriging procedures. Because it makes estimates for the average value within a block, block kriging is better suited for model evaluation than the practice of simply kriging to the center point of each grid cell. While the estimates of the overall spatial field are likely to be very similar, the standard error estimates yielded by block kriging are more accurate. This means greater precision in identifying grid cells whose modeled values are significantly different from what we expect to see based on observational data. When faced with the choice between kriging to the center of each grid cell vs. block kriging, the user should only choose the former if computer resources or time are limited, the area under consideration is large, and an estimate of the overall spatial field (no variability estimate) is all that is needed.

We note that block kriging, like most other classical kriging techniques, is subject to certain limitations. For instance, even though kriging depends on the estimation of the covariance structure of the spatial field, the standard errors do not include the potential error associated with this estimate. This structure is likely to be most difficult to estimate when the availability of monitoring data is limited. While Bayesian methods, such as those presented by Handcock and Stein (1993), Fuentes and Raftery (2005), and Swall and Davis (2006), can help address this problem, they are much more difficult to implement. More fundamentally, kriging

typically assumes stationarity, i.e. assumes that the spatial covariance structure is the same throughout the entire region of interest. This assumption is unlikely to hold true for long distances or across areas containing many different geological formations, and this limits the size of the region which can be addressed using these techniques. Also, it may be reasonable to closely investigate this assumption in regions which incorporate different land-use patterns and population densities, such as the urban/rural differences in the coastal and inland areas of the northeastern U.S. in the examples presented in Section 4.

DISCLAIMER: The United States Environmental Protection Agency through its Office of Research and Development funded and managed the research described here. It has been subjected to Agency administrative review and approved for publication.

References

Appel, K. W., Gilliland, A. B., Sarwar, G., Gilliam, R. C., 2007. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance Part I - Ozone. Atmospheric Environment 41, 9603–9615.

Byun, D., Schere, K. L., 2006. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Applied Mechanics Reviews 59, 51–77.

Cressie, N. A. C., 1993. Statistics for Spatial Data, Revised Edition. John Wiley and Sons, Inc.

Davis, J. M., Swall, J. L., 2006. An examination of the CMAQ simulations of the wet deposition of ammonium from a Bayesian perspective. Atmospheric Environment 40, 4562–4573.

Eder, B., Kang, D., Mathur, R., Yu, S., Schere, K., 2006. An operational evaluation of the Eta-CMAQ air quality forecast model. Atmospheric Environment 40, 4894–4905.

Eder, B., Yu, S., 2006. A performance evaluation of the 2004 release of Models-3 CMAQ. Atmospheric Environment 40, 4811–4824.

Fuentes, M., Guttorp, P., Challenor, P., 2003. Statistical assessment of numerical models. International Statistical Review 71, 201–221.

Fuentes, M., Raftery, A. E., 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61, 36–45.

Gelfand, A. E., Zhu, L., Carlin, B. P., 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.

Gotway, C. A., Young, L. J., 2002. Combining incompatible spatial data. Journal of the American Statistical Association 97, 632–648.

Grell, G. A., Dudhia, J., Stauffer, D. R., 1995. A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). NCAR Technical Note NCAR/TN-398+STR, National Center For Atmospheric Research, Boulder, CO.

Handcock, M. S., Stein, M. L., 1993. A Bayesian analysis of kriging. Technometrics 35, 403–410.

Isaaks, E. H., Srivastava, R. M., 1989. An Introduction to Applied Geostatistics. Oxford Universtiy Press.

Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30, 683–691.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Schlather, M., 2001. Simulation and analysis of random fields. R News 1 (2), 18–20.

Swall, J. L., Davis, J. M., 2006. A Bayesian statistical approach for the evaluation of CMAQ. Atmospheric Environment 40, 4883–4893.