Chapter 5

The Consequences of Interpolating or Calculating First: A Simple Case Study

This chapter marks the beginning of the part of the thesis oriented towards issues in pesticide leaching modelling. A synthetic case study is created to investigate the consequences of interpolating or calculating first in spatially distributed modelling.

5.1 Introduction

In this chapter, a synthetic, simple case study is constructed for the study of spatialisation and aggregation strategies in spatially distributed modelling. The first approach, *calculate alone* (CA), consists of the model application on point data followed by the aggregation of the results to the scale of the study area. Two further approaches are used to generate spatial output and differ by interpolating after or before the model run on point support (*calculate first, interpolate later*; CI vs. *interpolate first, calculate later*; IC).

Working on a simplified example is useful to eliminate constraints associated with real data sets or complex models. Moreover, validation is made possible and this allows to evaluate the different modelling strategies in an objective way.

The main objective of this chapter is to determine the influence of calculating or interpolating first when moving from point support data to spatially

aggregated block outputs. The results are confronted with the validation data set using map comparison statistics and cumulative density functions. Finally, the influence of model non-linearity is examined by testing the different methods with both a linear and a non-linear model.

5.2 Validation data set

In Matlab™ (version 7.0), **Z** was generated using a sequential simulation algorithm (BMElib; Christakos et al., 2002). **Z** is a zero-mean Gaussian distributed (variance = 3) spatial random field representing the 'reality' (i.e. validation data set), on a square area of 128×128 cells of size 2.5 (Figure 5.1). This scale of 2.5 will be further denoted as the 'point' resolution. The spatial structure of **Z** was defined by an exponential covariance model with a sill of 3 and a range of 80 (no nugget effect). When aggregated to cell size of 10, the spatial random field is said to be at the 'block' resolution (32×32 cells; Figure 5.2). The value for a block is the arithmetic average of the fine grid simulated values over the same block.

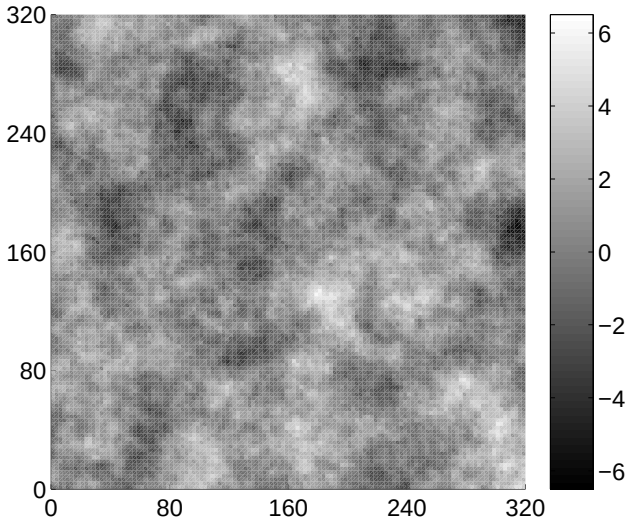


Figure 5.1: **Z** at the point resolution (cell size = 2.5). Exponential covariance model with parameters 3 (= sill) and 80 (= range).

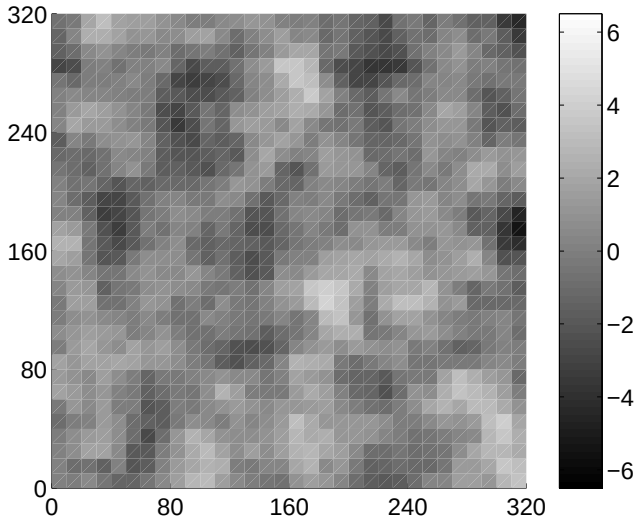


Figure 5.2: **Z** at the block resolution (cell size = 10), obtained from the spatial aggregation of Figure 5.1.

5.3 Models and methods

Figures 5.1 and 5.2 are used as the validation data, just as if they represented observations. Let assume that these observations can be estimated using either a linear or a non-linear model.

The linear model has been chosen as:

$$z_l = 2x_l - 5 \tag{5.1}$$

where z_l is the linear model output, which will be confronted to the validation data set **Z**; and x_l is the linear input data obtained by model inversion (cf. Eq. (5.3)).

The non-linear model has been chosen as:

$$z_{nl} = \exp(x_{nl}) - 10 \tag{5.2}$$

where z_{nl} is the non-linear model output, which will be confronted to the validation data set **Z**; and x_{nl} is the non-linear input data obtained by model inversion (cf. Eq. (5.4)).

If we consider that the system is perfectly known (i.e. model error is null for both models), we can derive the input data sets $\mathbf{X_l}$ and $\mathbf{X_{nl}}$ by model inversion on every pixel of the $\mathbf{Z}$ field at point resolution:

$$x_l = \frac{z_l + 5}{2} \tag{5.3}$$

$$x_{nl} = \ln(z_{nl} + 10) \tag{5.4}$$

Thus $\mathbf{Z}$ is the only random field generated; $\mathbf{X_l}$ and $\mathbf{X_{nl}}$ being derived by models inversion. Figures 5.3 and 5.4 display $\mathbf{X_l}$ and $\mathbf{X_{nl}}$, which represent fields of input data, upon which random samples will be collected in order to simulate a simplified modelling study: sampling → modelling/interpolation → validation.

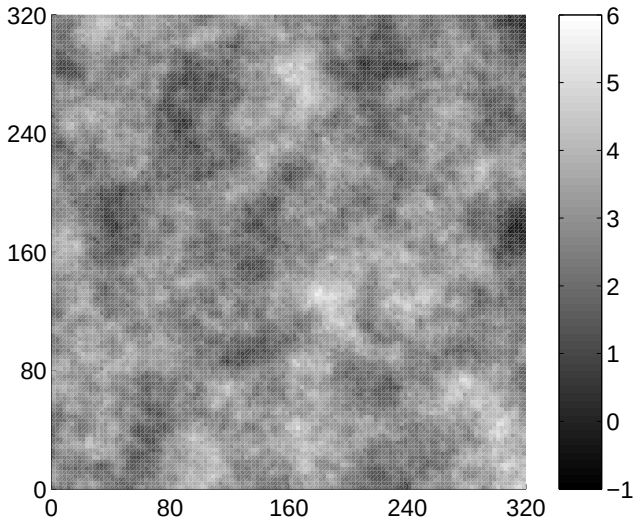


Figure 5.3: $\mathbf{X_l}$ at the point resolution (cell size = 2.5), obtained from the linear model inversion (Eq.5.3).

A number of 50 samples were located randomly in the study area (Figure 5.5). The number of samples was chosen to have a density sufficient to allow the derivation of an empirical semivariogram at a later stage. For the sake of the exercise, let us assume that these samples are the only available input data (e.g. soil profiles).

Three methods will be tested to transform this point information into block output.

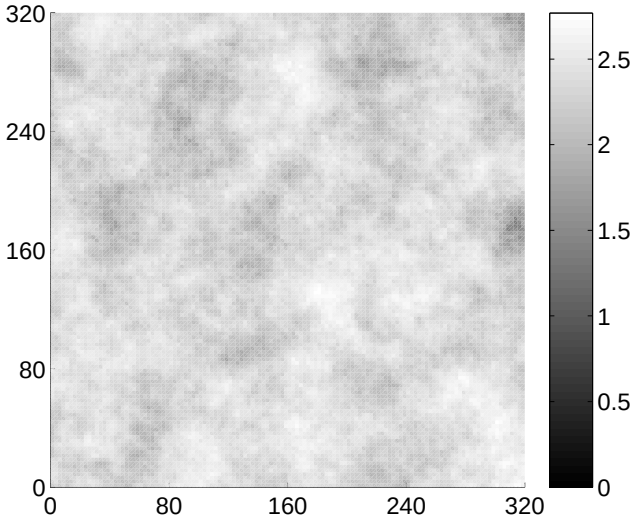


Figure 5.4: $\mathbf{X_{nl}}$ at the point resolution (cell size = 2.5), obtained from the non-linear model inversion (Eq.5.4).

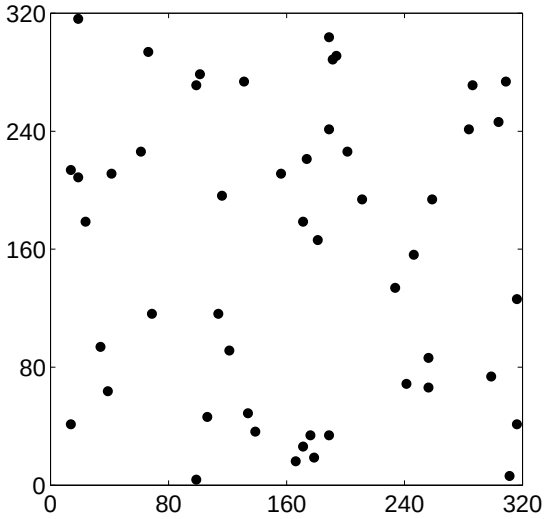


Figure 5.5: Location of the 50 samples collected in the study area.

The first is denoted as CA (for ***C****alculate* ***A****lone*) and is implemented as follows:

- Run the model on point support (50 soil profiles) and compute the output variable (z_l or z_{nl}).

This approach is not a change of support in the first sense, except if we consider the whole study area as being the block output.

The second approach is called CI (for ***C****alculate first,* ***I****nterpolate later*):

- Run the model on point support (50 soil profiles) and compute the output variable (z_l or z_{nl});
- Interpolate the output variable (using ordinary kriging) on a grid at point resolution, using (i) the original semivariogram model (CI1), or (ii) a semivariogram model fitted to the output variable (CI2);
- Aggregate the results at the block resolution.

The third approach is denoted as IC (for ***I****nterpolate first,* ***C****alculate later*):

- Interpolate the input variable (x_l or x_{nl}), using either (i) the transformation of the original semivariogram[1] (IC1), or (ii) a new fitted semivariogram (IC2);
- Run the model on point support;
- Aggregate the results at the block resolution.

The objective of this study is to investigate how model linearity or non-linearity interacts with kriging and the methodology for the change of support (CA, CI or IC). The outputs that will be considered are the maps and cumulative density functions (+ 95% confidence interval) at the point and block resolutions. The aggregation to the block resolution is a methodological step beyond the calculation/interpolation issue, but its interest is notably

[1] The semivariogram can be derived analytically for the linear model or by a numerical integration for the non-linear model.

to examine how the uncertainty is affected by the correlation of the points within the block prior to aggregation. Moreover, this aggregation allows the budget equations of Pontius et al. (2005) to be plotted over multiple resolutions (see next paragraphs) and thus to examine the extent and persistence of the agreement and disagreement due to location over increasing aggregation.

The comparison of output maps was performed using the statistical measures of Pontius et al. (2005). This method compares two maps that show (different) patterns of the same real variable, e.g. a simulation vs. a 'reference' map. The technique of Pontius et al. (2005) budgets the agreement and disagreement between the maps in terms of components of quantity (i.e. overall average) and location (i.e. spatial arrangement) of the variable. These budgets (see Eq. (5.5) to (5.10)) were computed using both the root mean square error (RMSE) and the mean absolute error (MAE). Let X_i and Y_i denote the reference and comparison values of the i^{th} pixel of the map (on a total of N pixels). For a given spatial resolution and assuming that all pixels have the same weight, the budget equations can be written as:

$$\text{Disagreement due to quantity (RMSE)} = \sqrt{\left[\frac{\sum_{i=1}^{N}(Y_i - X_i)}{N}\right]^2} \quad (5.5)$$

$$\text{Disagreement due to quantity (MAE)} = \left|\frac{\sum_{i=1}^{N}(Y_i - X_i)}{N}\right| \quad (5.6)$$

$$\text{Disagreement due to location (RMSE)} = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - X_i)^2}{N}} - \sqrt{\left[\frac{\sum_{i=1}^{N}(Y_i - X_i)}{N}\right]^2} \quad (5.7)$$

$$\text{Disagreement due to location (MAE)} = \frac{\sum_{i=1}^{N}|Y_i - X_i|}{N} - \left|\frac{\sum_{i=1}^{N}(Y_i - X_i)}{N}\right| \quad (5.8)$$

$$\text{Agreement due to location (RMSE)} = \sqrt{\frac{\sum_{i=1}^{N}\left(\overline{Y}-X_i\right)^2}{N}} - \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i-X_i\right)^2}{N}} \quad (5.9)$$

$$\text{Agreement due to location (MAE)} = \frac{\sum_{i=1}^{N}\left|\overline{Y}-X_i\right|}{N} - \frac{\sum_{i=1}^{N}\left|Y_i-X_i\right|}{N} \quad (5.10)$$

In Eq. (5.9) and (5.10), $\overline{Y}$ is the average of the variable Y. RMSE and MAE give the same disagreement due to quantity (Eq. (5.5) and (5.6) are the same), which is in fact the absolute difference between the average predicted value and the average observed value. This error measure is independent from resolution, because resolution modifies information of location, not information of quantity. The component of disagreement due to location (Eq. (5.7) and (5.8)) indicates how information of location changes with resolution. All components of information of location must ultimately become zero at the coarsest resolution (i.e. when the whole study area is just one aggregated block). When the disagreement due to location is plotted against the resolution (aggregation level), the rate at which the disagreement due to location shrinks indicates the distances at which the errors of location exist in the fine-resolution maps. If the disagreement due to location at a given resolution can be corrected by switching contiguous pixels, then it will be cleared at the next aggregation level. On the contrary, a disagreement of location that remains fairly stable over a certain resolution means that the disagreement can only be corrected by switching pixels over the corresponding distance. If the component of agreement due to location is null (Eq. (5.9) and (5.10)), this means that the 'reference' map would have been better approximated by an averaged uniform map.

The differences in shapes of the cumulative density functions (CDFs) were tested by an adaptation of the Kolmogorov-Smirnov (KS) test. This test assumes independence, which is not true for the spatially correlated data used in this study. The CDFs comparison was therefore made using KS in a Monte Carlo permutation test (Manly, 1997). In this methodology, the null hypothesis (H_0) is that the two distributions are identical; i.e. they are two samples from the same population. The distribution of the KS statistic

under H_0 is obtained by permuting the grouped observations in all possible ways ($N = 9999$ here) and with equal probabilities. This distribution is then used to calculate the significance of the test as:

$$p_{value} = \frac{1 + \#(\mathrm{KS}_i > \mathrm{KS}_0)}{1 + N} \tag{5.11}$$

where # counts the 'number of times', KS_i is the KS statistic of the i^{th} permutation and KS_0 is the value of the KS statistic for the original two samples. This statistical analysis of the CDFs is not spatially explicit and the test is liberal in the sense that it tends to reject the null hypothesis of equality of CDFs too often (Van Bodegom et al., 2002b).

These two comparison methods (budget equations and CDF comparison) will be applied at all resolution, from the point support to the coarsest resolution (i.e. one pixel for the whole surface). However, presentation of the results is limited to the point and block resolutions, for which the uncertainty is also computed.

5.4 Results

5.4.1 CA approach

In this study, because the system is perfectly known (no model error), the CA methodology will be only a measure of how well does the sampling represent the study area. The greater the number of samples, the better they are expected to be an approximation of the system behaviour for the study area. Yet, it must be reminded that this approach does not allow a true spatial representation of the results. Also, the CA method will yield the same results for both the linear and the non-linear models.

Figure 5.6 displays the cumulative density functions (CDFs) of $\mathbf{Z}$ at the point and block support, together with the CDF of the CA approach applied to the linear and non-linear models for a sample size of 50. The extent to which the CA CDF in Figure 5.6 (solid line) is close to the (dash-)dotted lines is only dependent on sample size and location. The effect of spatial aggregation on $\mathbf{Z}$ can be distinguished in the slight narrowing of

the block support CDF (dash-dotted line) compared to that of the point support (dotted line). However, extreme values are much more truncated in the CA approach. Depending on sample size, this can have some effect on the calculation of extreme percentiles. On Figure 5.6, percentile calculation does not seem to be strongly affected between the 5^{th} and 95^{th} percentile. Still, any interpretation would depend on the sample size and also on the confidence interval needed on the percentile calculation.

This confidence interval can be assessed using bootstrap. This method consists of resampling the 50 model outputs n times with replacement, assuming independence between the sampled values. In the present case, the 80^{th} percentile was 1.67. The distribution of the 80^{th} percentile was determined using bootstrap with $n = 500$ and the corresponding histogram is shown on Figure 5.7. The 90^{th} confidence interval was $[0.64; 2.19]$ (median $= 1.67$). This interval includes the 'true' 80^{th} percentile (1.40 at the point resolution). The assumption of independence is reasonable given that the samples were located at random.

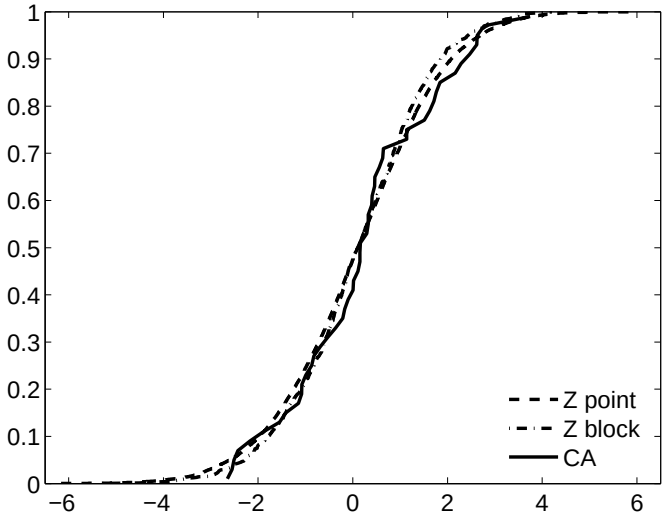


Figure 5.6: Cumulative density functions of $\mathbf{Z}$ (point and block supports) and the CA method applied to the linear or non-linear model. Results are identical for both the linear and non-linear models because there is no model error; i.e. the two models give the same output.

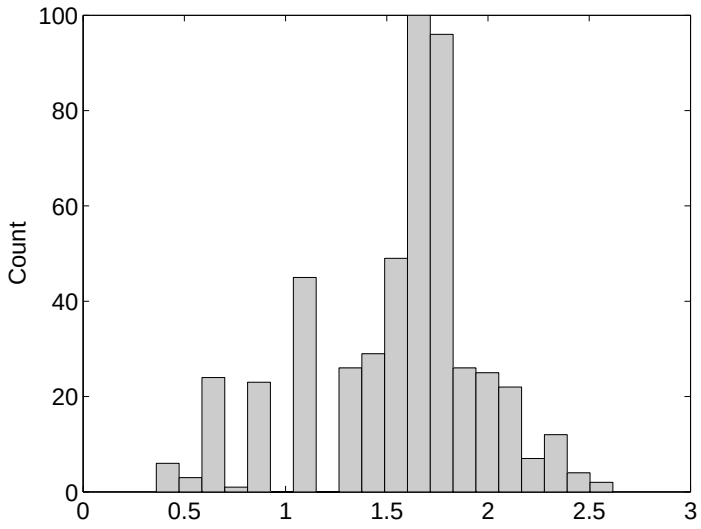


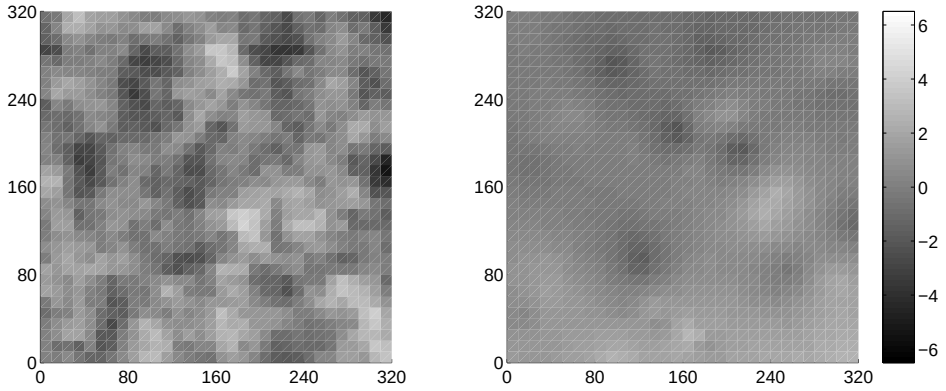
Figure 5.7: Histogram of the 80^{th} percentile of the CA method, caculated using a bootstrap procedure ($n = 500$).

5.4.2 CI approach

For the CI methodology, the same 50 samples are taken but this time, interpolation is performed after the model run on the 50 points. Here two cases are considered: (i) we 'know' the initial semivariogram to apply in the kriging algorithm (CI1); or (ii) we derive an empirical semivariogram based on the 50 output variables (CI2). In the first case, the quality of the final map will simply be a function of the sampling density because no error can be introduced in the semivariogram fitting. Figure 5.8 presents the results under this assumption on the block support. The mean error computed at the point (or block) resolution is equal to 0.0142. Figure 5.9 shows the results (identical for both the linear and non-linear models) of the budget equations. The disagreement due to quantity is equal to the mean error. Disagreement due to location is larger than the agreement due to location until an aggregation level of 32. Figure 5.10 presents the CDFs of the results at the block resolution, including the 95% confidence interval. The Generalised Least Squares methodology was applied to compute the block variance. This allowed to take the within-block covariance into account. The decrease in variability compared to the validation data is due to intrinsic regression characteristics in kriging. However, the aggregation

to the block support attenuates the difference between the CDFs because the extreme values of the validation data set are averaged. The results of the Monte Carlo permutation test (Manly, 1997) showed that the CDFs of CI1 and 'reality' were no longer statistically different from an aggregation level of 16 onwards. This means that at a coarser resolution than 16×16, the two approaches produced statistically similar results.

A word of caution is given here concerning Figure 5.10 (and subsequent similar figures). Theoretically, it is not correct to represent the CDF of kriging estimates, as they do not belong to the same distribution. However, one aim of this exercise is to show how CI or IC can affect the distribution of a variable in a spatially distributed case, and also to see the effect on indicators such as upper percentiles used in pesticide registration procedures (see for example chapter 8). Thus, plotting the CDF of the kriging estimates allows the consequences of the scaling approaches tested in this chapter and used in practical applications to be captured. It is also acknowledged that the CDF of the kriging estimates should not be expected to compare well with the CDF of the 'validation' values. However, Figure 5.10 provides an illustration of the regression towards the mean induced by the kriging procedure.



(a) (b)

Figure 5.8: Block support of the (a) 'validation data set' **Z** and (b) CI1 approach for both the linear and non-linear models, assuming that the semi-variogram of the output variable **Z** is perfectly known. Results are identical for both the linear and non-linear models because there is no model error; i.e. the two models give the same output.

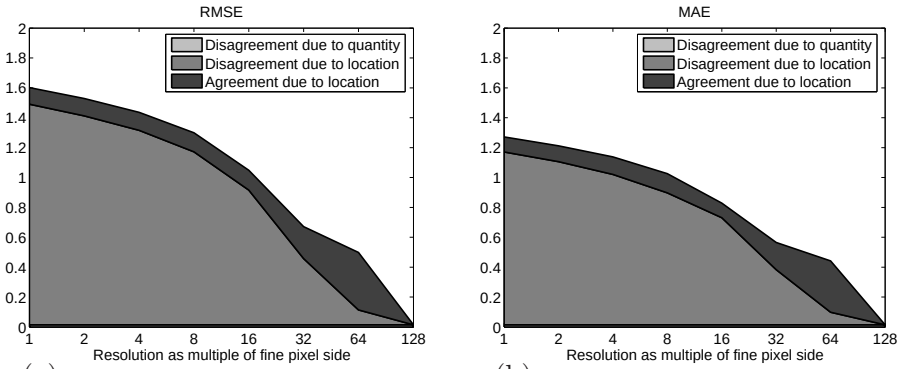


Figure 5.9: (a) Root Mean Square Error (RMSE) and (b) Mean Absolute Error (MAE) for the CI1 approach. Note that the disagreement due to quantity is almost null.

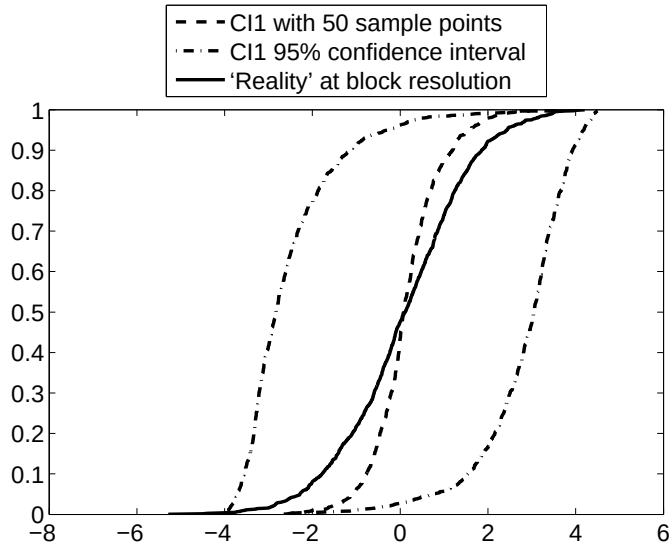


Figure 5.10: Cumulative density function and confidence interval (95%) for the output at block resolution of the CI1 approach for both the linear and non-linear models (dashed line), assuming that the semivariogram of the output variable $\mathbf{Z}$ is perfectly known. The CDF of the 'true' data is displayed by the solid line.

However, in the second case (CI2), the sampling density is also very important, as it determines the accuracy to which the semivariogram model can be estimated. With the sampling scheme of Figure 5.5, the exponential semivariogram model fitted to the data had the following parameters: sill = 2.28 and range = 79.82. These parameters have to be compared with the 'true' parameters of 3 and 80. The difference in spatial range is negligible. However, this seemed to be a coincidence, as shown in Figure 5.11. In this figure, it can be noticed that depending on sample size, the 'true' range of 80 is correctly approximated within a ±20 interval.

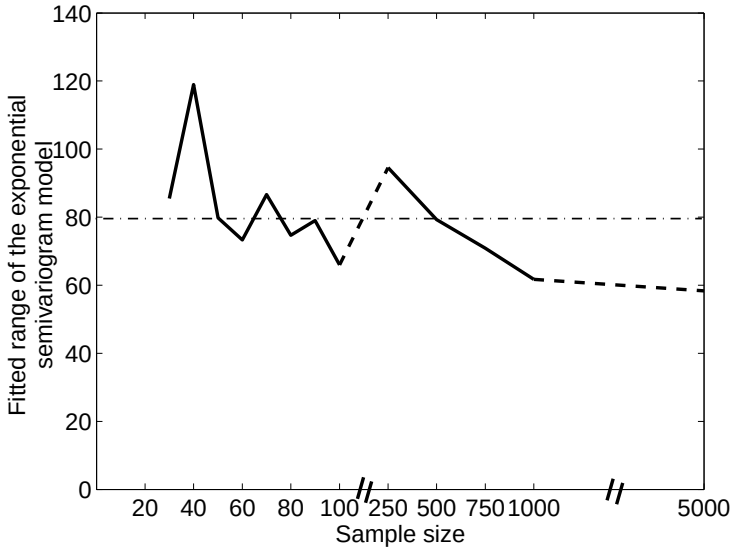
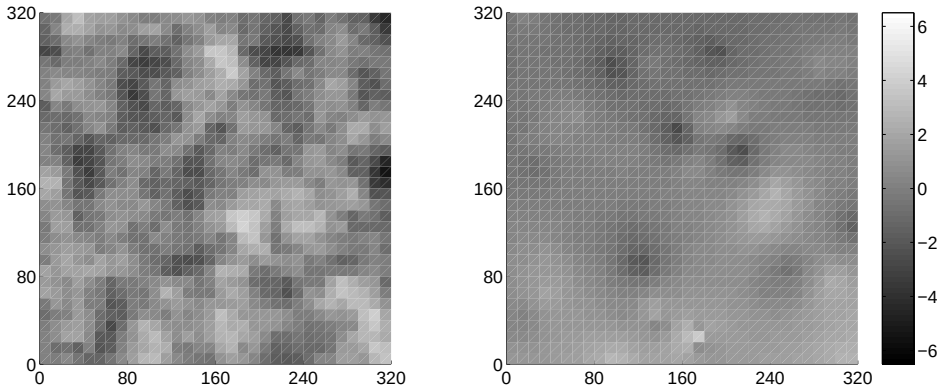


Figure 5.11: Spatial correlation range fitted to the point data in the CI2 approach, given as a function of the sample size. The dash-dotteded line marks the 'true' correlation range (= 80) that was used to generate the original data set.

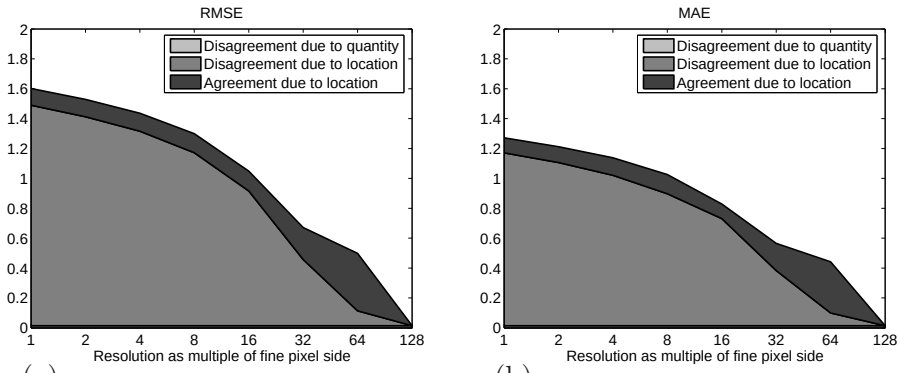
Once the parameters of the semivariogram model have been estimated, the next step is to interpolate the point simulation (at the point resolution), and then to aggregate the results at the block support. Figures 5.12 to 5.14 present the results for the case where 50 samples were used. The results are very similar between CI1 and CI2, since the fitted semivariogram is almost identical to the original one (only the sill is smaller, but this had no particular incidence on the results). The mean error (or disagreement due to quantity) computed at the point (or block) resolution for CI2 is equal to

0.0141, which is equal to the mean error of CI1.



(a) (b)

Figure 5.12: Block support of the (a) 'validation data set' **Z** and (b) CI2 approach for both the linear and non-linear models, fitting an exponential semivariogram model to a sample of the output variable.



(a) (b)

Figure 5.13: (a) Root Mean Square Error (RMSE) and (b) Mean Absolute Error (MAE) for the CI2 approach.

At this point, another seed number in the random allocation of the 50 samples was used, to see to what extent the fitted semivariogram range is sensitive to the seed number. For exemple, seed number 5 instead of 0 led to a semivariogram fit with the sill and range equal to 2.41 and 58.40 respectively. Consequently, the mean error for that simulation was equal to 0.2022. The map at block resolution is displayed in Figure 5.15. The results of the RMSE and MAE budget equations for this CI2 simulation

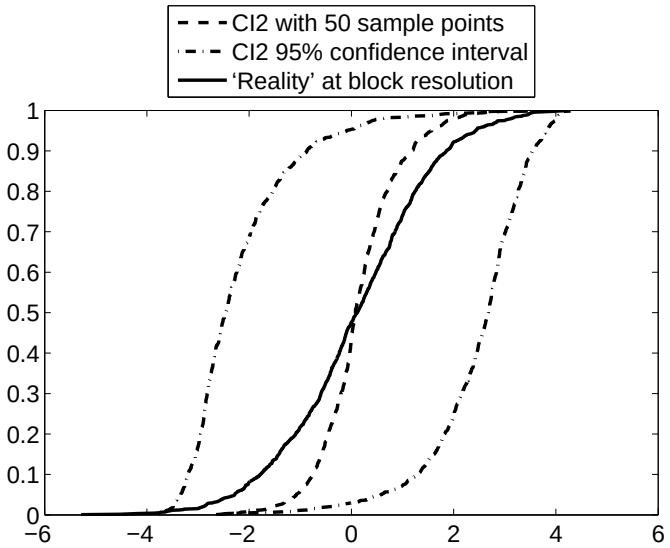


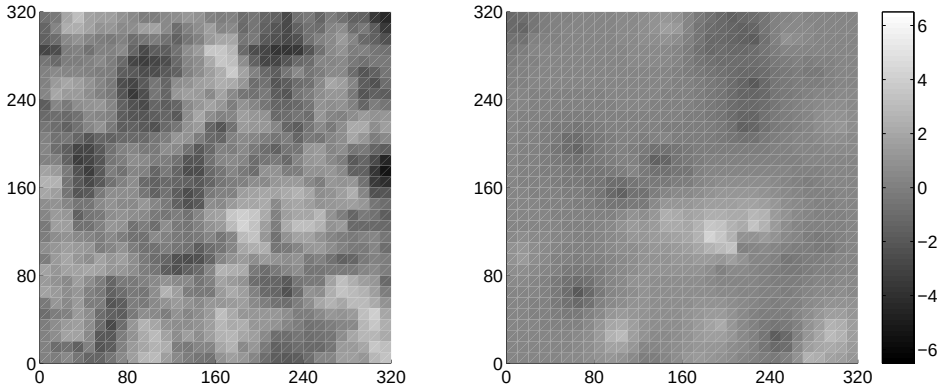
Figure 5.14: Cumulative density function and confidence interval (95%) for the output at block resolution of the CI2 approach for both the linear and non-linear models (dashed line), fitting an exponential semivariogram model to the output variable. The CDF of the 'true' data is displayed by the solid line.

with a different seed number are displayed in Figure 5.16. Similarly to the previous results, the Monte Carlo permutation test (Manly, 1997) showed that the CDFs of CI1 and 'reality' were no longer statistically different from an aggregation level of 16 onwards.

Figure 5.16 is very similar to Figure 5.13, except for the amount of disagreement due to quantity. This suggests that the spatial agreement of the CI simulations is relatively independent of sample location. The agreement due to location is never null, meaning that the CI approach does better than would an average uniform map. However, the disagreement due to location is significantly higher than the agreement due to location at all resolutions finer than 32 times the point resolution.

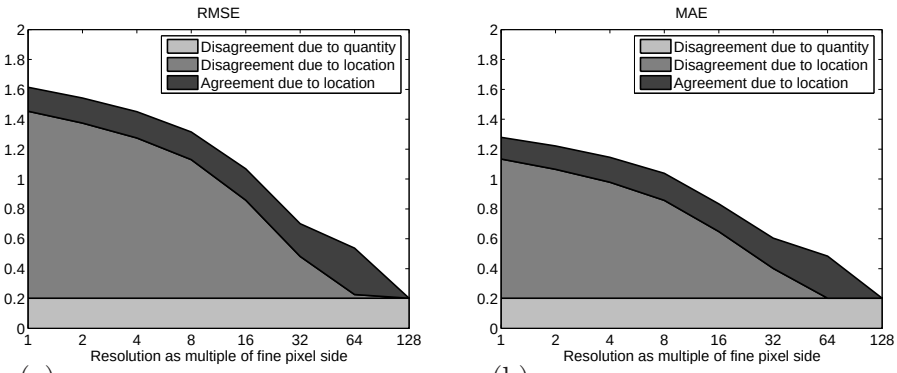
5.4.3 IC approach

From a theoretical viewpoint, the IC approach should be equivalent to the CI approach in the case of a linear model, because only linear operations are involved. The results are identical when the 'true' semivariogram model



(a) (b)

Figure 5.15: Block support of the (a) 'validation data set' **Z** and (b) CI2 approach for both the linear and non-linear models, fitting an exponential semivariogram model to a sample of the output variable, with a different seed number.



(a) (b)

Figure 5.16: (a) Root Mean Square Error (RMSE) and (b) Mean Absolute Error (MAE) for the CI2 approach with a different seed number.

is known (IC1). The uncertainty on the output CDF was analysed using conditional simulations in the interpolation step. Figure 5.17 shows the CDFs of the median, 5^{th} and 95^{th} percentiles of 20 simulations (in the case of the linear model, with known semivariogram).

Figure 5.18 shows the output map at the block resolution, in the case where the semivariogram model is not known (i.e. fit to a sample of size 50). The correlation range fitted to the data is not affected by the methodology chosen (CI or IC), and hence the results are the same. Indeed, Figure 5.18(b) is exactly the same as Figure 5.12(b).

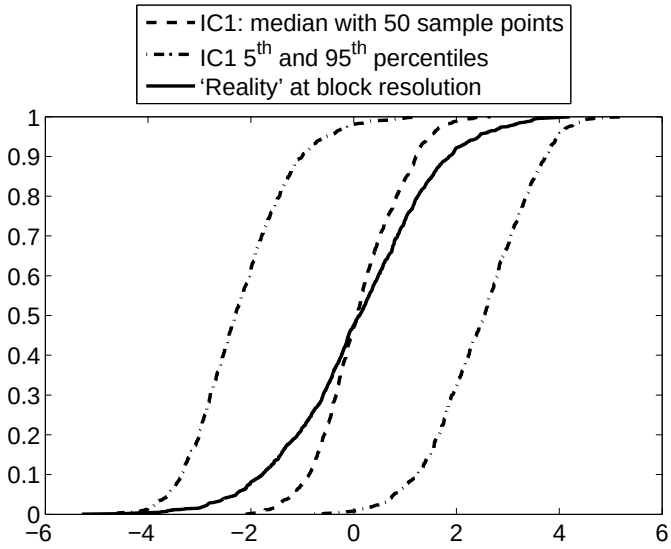
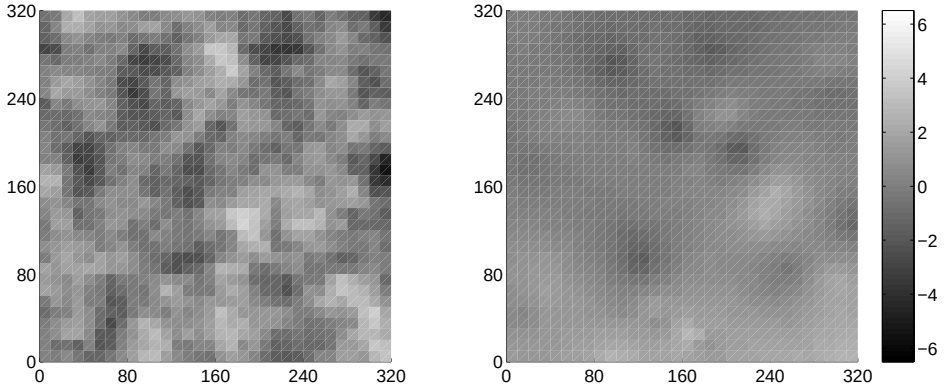


Figure 5.17: Cumulative density function of the median, 5^{th} and 95^{th} percentiles of 20 simulations for the IC1 approach at block resolution with the linear model (dashed line), knowing the exact semivariogram model. The CDF of the 'true' data is displayed by the solid line.

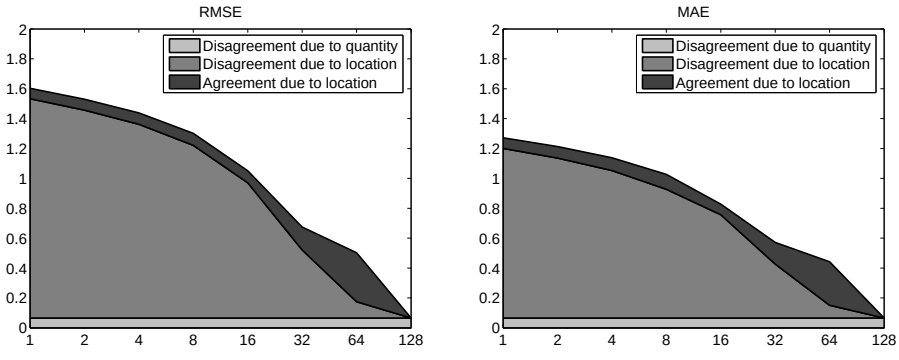
When using the non-linear model, the IC approach is no longer equivalent to the CI approach. However, although $\mathbf{X_{nl}}$ resulted from a logarithm transformation of $\mathbf{Z}$ (Eq. (5.4)), it still follows a normal distribution (Kolmogorov-Smirnov test; p-value $= 0.84$). Therefore, ordinary kriging can be used in the interpolation step even though a bias might be introduced here.

For the 'IC1' approach (i.e. the semivariogram model is known), one has to derive the covariance function from the original model. Eq. (5.12) is



(a) (b)

Figure 5.18: Block support of the (a) 'validation data set' **Z** and (b) IC2 approach for the linear model, fitting an exponential semivariogram model to a sample of the input variable $\mathbf{X_l}$.



(a) (b)

Figure 5.19: (a) Root Mean Square Error (RMSE) and (b) Mean Absolute Error (MAE) for the IC2 approach with the linear model.

known from the original model:

$$\mathrm{cov}[Z_i, Z_j] = \mathrm{E}[Z_i Z_j] - \mathrm{E}^2[Z] \tag{5.12}$$

where Z is the Gaussian variable, and Z_i and Z_j are the values of the variable in two points separated by a given distance.

To derive the covariance function of a non-linear transformation of the variable Z, Eq. (5.13) has to be calculated for every distance class:

$$\begin{aligned} \mathrm{cov}[\varphi(Z_i), \varphi(Z_j)] &= \mathrm{E}[\varphi(Z_i)\,\varphi(Z_j)] - \mathrm{E}^2[\varphi(Z)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(Z_i)\,\varphi(Z_j)\,f(Z_i, Z_j)\,dz_i dz_j \\ &\quad - \left(\int_{-\infty}^{\infty} \varphi(Z)\,f(Z)\,dz\right)^2 \end{aligned} \tag{5.13}$$

where φ is the transformation function; f is the probability density function. The covariance function obtained by this transformation was approximated by an exponential model having a sill of 20.28 and a range of 17.98 (Figure 5.20).

Figures 5.21(a) and 5.22 show the results for the IC1 approach, with the non-linear model described by Eq. (5.2) and in the case where the covariance function is derived from the original one using Eq. (5.13). The block support of the CI1 approach is shown in Figure 5.21(b) to allow comparison with the IC1 approach. The mean error of the latter is equal to 0.0927, which is higher than the mean error in the CI1 approach. Disagreement and agreement due to location are slightly lower for IC1 than for CI1 at all aggregation levels.

Figure 5.23 computes the difference between the CI1 and IC1 approaches, for the non-linear model on point and block supports. The differences range between -1.34 and $+1.63$ for the point support, and between -1.22 and $+1.57$ for the block support. It can be seen that the areas of highest differences logically occur in the neighborhood of the samples used in the interpolation. This results from the change of semivariogram range (from 80 to 17.98 for CI1 and IC1, respectively); which in turn explains why the differences are limited in the very close neighborhood of the interpolation points, and gradually increase in the trasition between the respective semivariogram ranges (e.g. note the small grey surface surrounded by white at

the coordinates [240, 160] in Figure 5.23(a)).

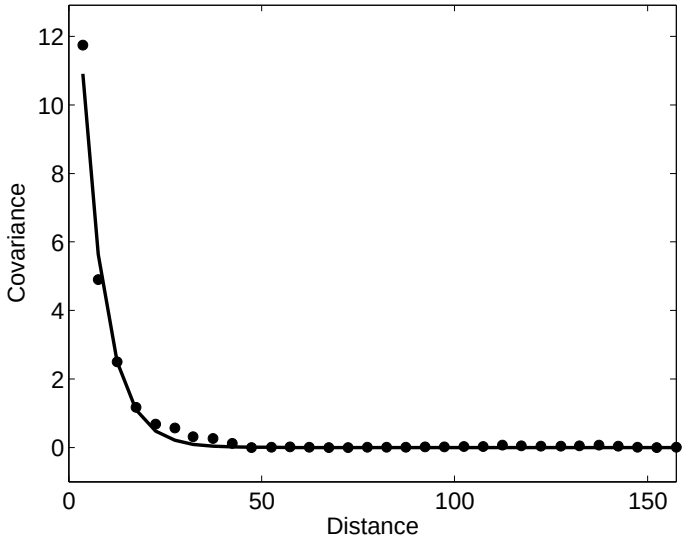
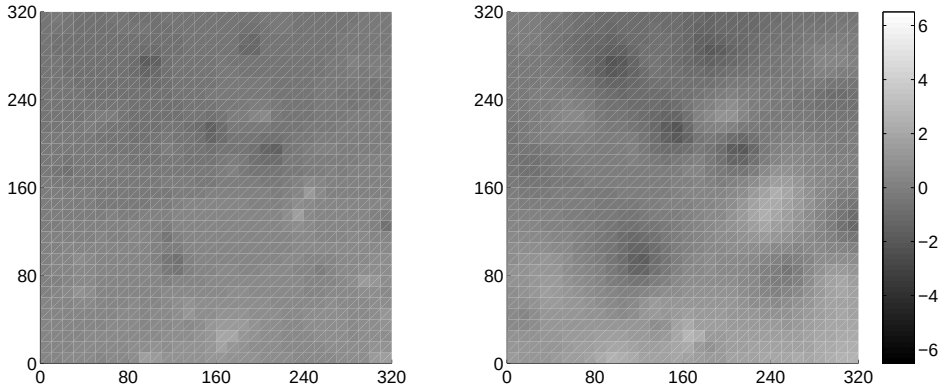


Figure 5.20: Covariance function fit to the non-linear transformation of the original covariance function, as described by Eq. (5.13).

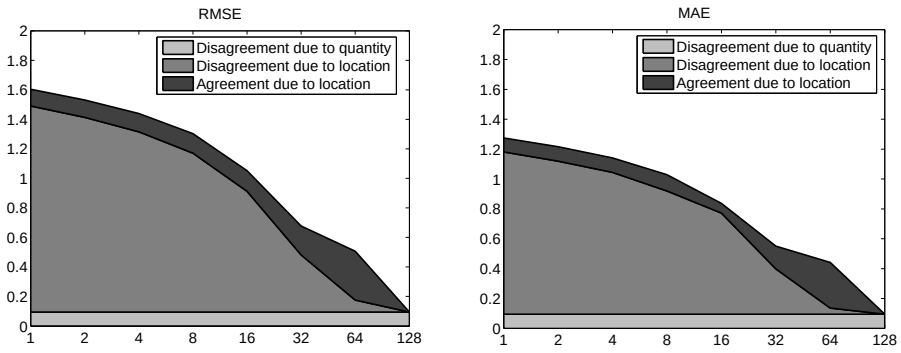
Figure 5.24 shows the CDFs of the median and 5^{th} percentile of 20 simulations for the IC1 approach with the non-linear model. Applying the non-linear model after the conditional simulations produced extremely high values, and the 95^{th} percentile CDF is therefore not visible in this figure because it extends over several orders of magnitude. This results from the bias caused by the application of ordinary kriging to the transformed $\mathbf{X_{nl}}$, although the samples were found to follow a normal distribution.

In the IC2 approach with the non-linear model, an exponential semivariogram was fit to the sample data with a sill of 0.0228 and a range of 63.92. The range is larger than in the IC1 approach, and thus the results are expected to be closer to the CI approach than IC1 was. Figure 5.25 presents the maps obtained with the non-linear model in the IC1 and IC2 approaches. The spatial patterns are identical (i.e. agreement and disagreement due to location are similar), but Figure 5.25(b) is more contrasted than Figure 5.25(a), due to the difference in range of the semivariogram models used in each case.



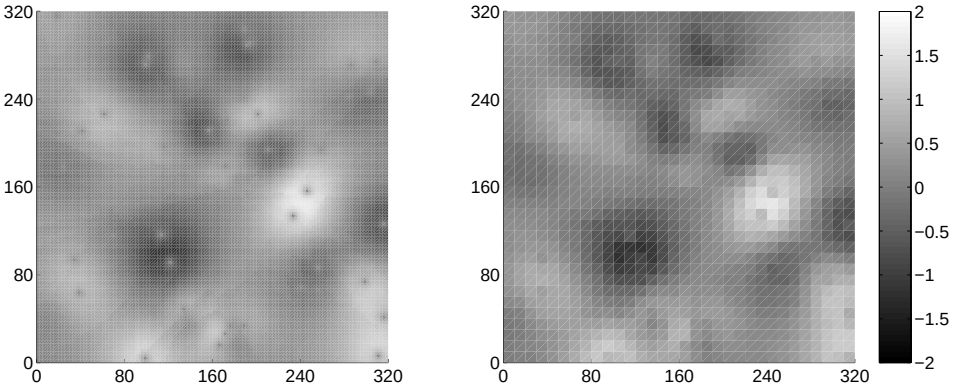
(a) (b)

Figure 5.21: Block support of the (a) IC1 approach for the non-linear model, numerically deriving the covariance function from the original model for each distance class; and (b) CI1 approach, knowing the exact semivariogram model.



(a) (b)

Figure 5.22: (a) Root Mean Square Error (RMSE) and (b) Mean Absolute Error (MAE) for the IC1 approach with the non-linear model.



(a) (b)

Figure 5.23: Difference between the CI and IC approaches (CI−IC) for the non-linear model, on the (a) point and (b) block supports.

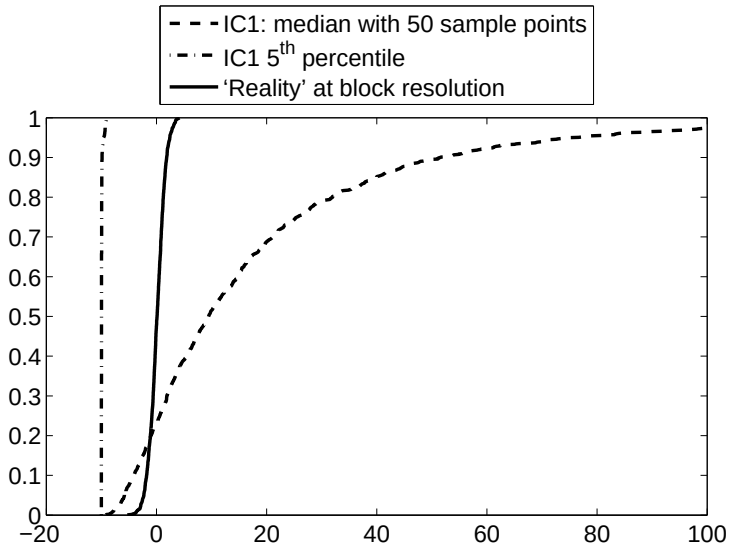
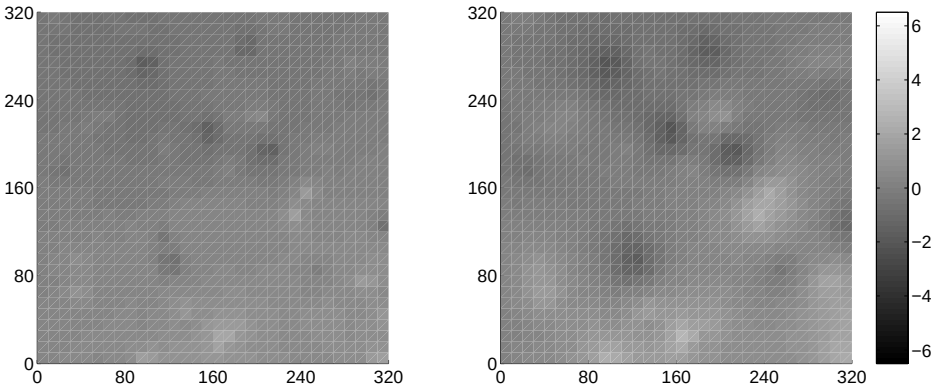


Figure 5.24: Cumulative density function of the 5^{th} percentile and median of 20 simulations for the IC1 approach at the block resolution with the non-linear model (dashed line), deriving a covariance function from the original model. The 95^{th} percentile curve is not visible on this plot because it starts from about 1000 on the x-axis and extends over several orders of magnitude. The CDF of the 'true' data is displayed by the solid line.



(a) (b)

Figure 5.25: Block support of the (a) IC1 approach for the non-linear model, numerically deriving the covariance function from the original model for each distance class; and (b) IC2 approach for the non-linear model, fitting an exponential semivariogram model to a sample of the input variable $\mathbf{X_{nl}}$.

5.5 Discussion and conclusions

This study illustrated how the CA approach can be implemented to assess the distribution of a model output at a certain scale. The advantage of this approach over the CI and IC methods is that no regression towards the mean is observed in the absence of an interpolation step. The disadvantage, however, is that the output is not spatially distributed. Assuming independence between the point samples, a standard bootstrap procedure can be used to evaluate the uncertainty around statistical indicators such as percentiles of the output distribution.

In the case of a linear model, the CI and IC approaches proved to be exactly the same, as expected from a theoretical viewpoint. The accuracy of model simulations is thus only dependent on sampling density (given that model error was considered to be null in the present case).

For the non-linear model, the CI and IC approaches produced different results. The differences were greater in the cases where the semivariogram models are known or derived from the original one, than in the case where a semivariogram is fitted to the data. Considerable differences were observed in the computation of the confidence intervals of the CDF (i.e. when the

IC approach is performed with conditional simulations). In this case, the bias introduced by the application of ordinary kriging to the transformed $\mathbf{X_{nl}}$ led to a complete shift of the output CDFs compared to the validation data. However, we chose to apply ordinary kriging because this reflects the method that would be used in real applications, if the samples follow a normal distribution as it was the case here.

In this study, we did not investigate the effect of sampling (location and quantity). Changes in the sampling pattern or density are likely to affect the results but not the conclusions about the different approaches relative to one another. Some results (e.g. change of seed number) suggested that the influence of sampling could be much greater than the influence of choosing between CI and IC. Overall, the differences between these two approaches were rather limited for the first moment of the results, even for the non-linear model. However, the comparison between the different approaches was made for one possible choice of a non-linear model, so the conclusions do not necessarily extend to other possible models. Besides, only one variable was included in the non-linear model, while for real case studies several input variables (possibly correlated) are typically involved and this could further affect the results.

One may argue that the CI approach should be preferred because it 'fills the missing information' in last resort. Furthemore, the risk that the IC method produces bias, such as in the conditionalised simulations (rather than kriging) reinforces this position. However, this study also showed that the differences observed between CI and IC for the non-linear model do not seem to concern the spatial pattern of the simulations. This means that the interpretation of the output maps based on a relative spatial comparison would not be affected. This conclusion is important in the context of relative groundwater vulnerability assessment.