Acceptable Ergodic Fluctuations and Simulation of Skewed Distributions

O. Leuangthong, J.A. McLennan and C.V. Deutsch
Dept. of Civil & Environmental Engineering, University of Alberta, Edmonton, AB Canada

ABSTRACT: Statistical fluctuations are an important part of stochastic simulation; however, the fluctuations should be reasonable and unbiased. Gaussian Simulation produces simulated values that are approximately standard normal in expected value. Minor fluctuations from a zero mean and unit variance are expected. The fluctuations will be larger when the range of correlation is large with respect to size of the domain. These fluctuations do not cause any bias in the back transformed realizations if the fluctuations are theoretically correct, that is, in keeping with the multivariate Gaussian random function model. The fluctuations may be exaggerated when certain implementation decisions are taken (ordinary kriging with a too small search neighborhood) or when the conditioning data do not follow the multivariate Gaussian model; in this case, the back transformation to original units may induce a bias in the mean in original units. This is particularly true for skewed distributions commonly encountered for geological variables.

Geostatistical simulation is becoming widely used to assess risk in resource assessment. The ability to construct multiple realizations is the key feature that permits uncertainty assessment and the transfer of joint uncertainty through planning. Gaussian simulation is the simplest robust simulation approach. The practitioner should take care with many implementation decisions and checking steps. This paper documents the consequences of statistical fluctuations, poor implementation choices and the non-linear transformation required in conventional Gaussian simulation. The problems are explained and some practical solutions are proposed.

1 INTRODUCTION

Gaussian simulation is being used increasingly in the mineral resource industry to construct models of heterogeneity and uncertainty. Simulation proceeds within deemed homogeneous rock types or geological units. The spatial extent of each rock type is modeled deterministically or stochastically before the simulation of continuous variables with Gaussian simulation. We are concerned only with the Gaussian simulation of continuous variables in this paper. The rock types and continuous variable realizations are merged in one of two methods (1) the continuous variable is simulated over the entire area independently of the rock type and then the large realizations are clipped according to the limits of the rock type model, or (2) the continuous variable is simulated only within the rock type. There is no theoretical difference in these approaches, but many practitioners prefer the first approach because it is more flexible and the input statistics are reproduced better. The fact that the input statistics are reproduced better is somewhat artificial since the large realization is clipped before being used for any heterogeneity or uncertainty assessment. Statistical or ergodic fluctuations are expected for anything less than an infinite domain.

The probability distribution of geologic variables is rarely Gaussian. Most variables are mass or volume concentrations that are bounded between 0 and 100%. Variables in small concentration are positively skewed and variables in large concentration tend to be negatively skewed. The original Z variable is transformed to a standard Gaussian variable prior to Gaussian simulation:

$$y = G^{-1}\left(F_z\left(z\right)\right) \tag{1}$$

where $F_Z(z)$ is the distribution of the original Z distribution. This distribution should be established carefully with declustering and/or debiasing as appropriate. The distribution $F_Z(z)$ and its inverse are tabulated functions from the available data. The $G^{-1}(\bullet)$ notation represents the inverse of the cumulative Gaussian distribution. Although there is no analytical solution to $G(\bullet)$ or $G^{-1}(\bullet)$, excellent polynomial approximations exist. Simulation proceeds in the Y-Gaussian units and the values are back transformed:

$$z = F^{-1}\left(G\left(y\right)\right) \tag{2}$$

It is important to note that the distribution of back transformed z-values need not match the initial reference distribution. The reference distribution $F_Z(z)$ is only reproduced exactly when the simulated y-values are exactly Gaussian with a mean of zero and a variance of one. As suggested above, the smaller the domain (rock type) the more statistical fluctuations we expect. Many practitioners associate these statistical fluctuations to uncertainty in the geologic variable, which is rea-

sonable. There is additional uncertainty due to uncertainty in the input distribution and other statistical parameters; however, we are concerned only with these statistical or ergodic fluctuations.

Some theoreticians and practitioners believe that the target distribution $F_Z(z)$ is reproduced exactly because of the transform/back transform. That is simply not true. The more the simulated *Y* values depart from a standard normal distribution, the more the back transformed *Z* values depart from the target distribution. A post-processing transformation could be considered to enforce the target distribution $F_Z(z)$ using, for example, the `trans` program of GSLIB (Journel & Xu, 1994); however, the fluctuations in the global distribution are often an important aspect of uncertainty. The aim of this paper is to understand those fluctuations and to assess the risk for bias.

2 ERGODIC FLUCTUATIONS

Simulation from a multivariate Gaussian distribution is often implemented by a sequential method. Although specific implementation choices are important, considering spectral or matrix simulation methods would lead to very similar results. The fluctuations described here are not limited to the sequential algorithm. The target reference distribution inside all Gaussian simulation implementations is the standard normal or Gaussian distribution with a mean of zero and a variance of one. Simulated realizations over a finite domain, however, would not exactly reproduce a mean of zero and a variance of one because of statistical fluctuations. The statistical fluctuations due to a finite domain size are sometimes referred to as *ergodic fluctuations*. In practice, *ergodicity* refers to how large the domain is relative to the range of correlation. The fluctuations are due to a lack of ergodicity, that is, the domain is small relative to the range of correlation. These fluctuations are significant.

Consider a stationary Gaussian random function (RF) representing a spatially correlated distribution of *N* grid nodes. The *N* grid nodes need not represent a contiguous 2 or 3-D array; the values could be clipped by geological boundaries or any other arbitrary limits. A large number (*L*) of realizations are simulated at each of the *N* grid node values. We could represent this set of realizations as:

$$\left\{ y_i^{(l)}, i = 1, ..., N; l = 1, ..., L \right\} \tag{3}$$

Where the superscript *(l)* denotes the realization number and the subscript *i* denotes the grid node index. The mean and variance of the realizations could be written:

$$\left.\begin{aligned} \overline{y}^{(l)} &= \frac{1}{N}\sum_{i=1}^{N} y_i^{(l)} \\ \sigma^{2,(l)} &= \frac{1}{N}\sum_{i=1}^{N} y_i^{(l)2} - \left[\frac{1}{N}\sum_{i=1}^{N} y_i^{(l)}\right]^2 \end{aligned}\right\} \; l = 1, ..., L \tag{4}$$

We equally weight these statistics because each grid cell is assumed to represent the same volume. The sampling distributions of the mean and variance in Equation 4 are of interest to us.

When the realizations are unconditional, the mean of the means and the variance of the means can be calculated fairly straightforwardly:

$$\begin{aligned} E\{\overline{y}\} &= \frac{1}{N}\sum_{i=1}^{N} E\{y_i\} = 0 \\ Var\{\overline{y}\} &= E\{\overline{y}^2\} - E\{\overline{y}\}^2 \\ &= E\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j\right\} \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} Cov\{y_i, y_j\} \end{aligned} \tag{5}$$

This is derived from simple linear algebra and the decision of stationarity. The covariance between the i^{th} value and itself is, by definition, the variance, which is one in standard Gaussian units; therefore, we can rewrite the variance in two parts:

$$Var\{\overline{y}\} = \frac{1}{N} + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} Cov\{y_i, y_j\} \tag{6}$$

The first *1/N* part is the result of classical statistics and the second part is due to the spatial correlation between the *N* grid nodes. In general the correlation is positive and the variance of the average is greater than the *1/N* predicted from independent data. The average covariance between all N^2 pairs of points has been used for many years in geostatistics. The distribution of the means has mean and variance values given in Equation 5, which are easily predicted since we know the variogram/covariance model used in the construction of the realizations. Moreover, the shape of the distribution is likely to be Gaussian since the data are Gaussian, the mean is a linear combination and a linear combination of Gaussian variables is also Gaussian.

The mean of the variances and the variance of the variances is more complex. In fact, we can work out an estimate for the expected value of the variances in the same way that we derived the result in Equation 5:

$$\begin{aligned} E\{\sigma^2\} &= \frac{1}{N}\sum_{i=1}^{N} E\{y_i^2\} - E\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j\right\} \\ &= 1 - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} Cov\{y_i, y_j\} \\ Var\{\sigma^2\} &= \text{a complex fourth order statistic} \end{aligned} \tag{7}$$

It would appear that the mean variance is closely related to the variance of the mean values (see Equations 5/6); however, we have not been able to verify this equation with simulation results. Alternative simulation algorithms have been considered and the calculated variance in Equation 7 appears to underestimate the mean variance in practice. The fourth order statistic for the variance is complex. It can be worked out and other (geo)statisticians may have done so in presence of spatial correlation. We suspect that this has been worked out by Matheron with an approximation in terms of the covariance function. In any case, the distribution of the variances has a mean given in Equation 7 and the variance could probably be worked out. The shape of the distribution, however, is ***not*** Gaussian. In general, the square of a Gaussian statistic follows

a gamma distribution (which is closely related to the Chi-square and Wishart distributions) (Johnson et al. 1994, Johnson & Wichern 1998). The shape of the distribution of the standard deviation (which cannot go negative) is also skewed and is fit by another type of gamma distribution.

There is a wide variety of notation and nomenclature used for the gamma distribution. The following is from the NIST Engineering and Statistics Handbook available on the web (see reference list):

$$f(x) = \frac{\left(\frac{x-\mu}{\beta}\right)^{\gamma-1} \exp\left(-\frac{x-\mu}{\beta}\right)}{\beta\,\Gamma(\gamma)} \qquad (8)$$

$$\text{where} \quad \Gamma(\gamma) = \int_0^\infty t^{a-1} e^{-t}\, dt$$

where μ is a location parameter (the minimum value), β is a scale or spread parameter and γ is a shape parameter. The $\Gamma(\gamma)$ function is known as the incomplete gamma function and there is no analytical solution. We used the polynomial approximation available in Numerical Recipes (Press et al. 2002). A γ value of about 2 is appropriate. The μ value is close to 1 for large domains and decreases as the ergodic fluctuations increase (domain size decreasing or range of correlation increasing). The β value is small for large domains and increases as the ergodic fluctuations increase.

Although the analytical solution to the sampling distribution of the variance is not readily available, we can establish the distributions by fitting the three parameter gamma distribution from simulated values. An iterative procedure will be used here. Fitting the distributions is somewhat unsatisfactory because we would have difficulty predicting the ergodic fluctuations in the variance ahead of time; however, we would have difficult with such predictions anyway because of conditioning data. The fluctuations in presence of conditioning data would be significantly more complex than the equations presented above. The numerical approach is quite useful.

2.1 *Small Example*

A small simulation study will be undertaken to illustrate the results described above. One thousand realizations of a 50 by 50 grid were generated for three cases: an isotropic range of 10 units, an isotropic range of 25 units and an isotropic range of 50 units. In all cases, a nugget effect of 0.1 was considered and a spherical structure with sill component of 0.9. The `sgsim` program from GSLIB (Deutsch & Journel 1998) was used for simulation. The search was set to the variogram range and a large number of previously simulated values were chosen (24) to ensure reasonable results. As required by theory, simple kriging was used in the simulation.

There is always a question of whether the results we see are due to true sampling of the multivariate distribution or to particular implementation choices. The `lusim` program from GSLIB was also considered to verify that the results are the same. The upfront matrix setup for `lusim` was significant, but both `sgsim` and `lusim` required about the same amount of computer time. The ergodic fluctuations from `sgsim` and `lusim` are virtually identical. The summary statistics of ergodic fluctuations shown in Figure 1 are the same within $1/10^{th}$ of 1 percent. The implementation of LU and SGS are radically different; thus, we are confident that the results are representative samples from a multivariate Gaussian distribution.

Figure 1 shows some results. The top row of Figure 1 shows the first realization for a visual appreciation of the scale of variability. The second row shows histograms of the mean values. The frequency scale changes, but the limits of -1.5 to 1.5 remain the same on all three histograms. As expected, the variability increases as the range of correlation increases (from left to right). The blue line is the expected mean of 0.0 and the red line is the expected Gaussian distribution for the mean values. The variance values for the expected Gaussian distributions were calculated as in Equation 5 (using the GSLIB-like `gammabar` program). The observed variance values match very closely to those predicted by theory (we expect standard deviations of 0.1364, 0.3170, and 0.5475).

The third row of Figure 1 shows histograms of the variance values. The blue line is the ergodic limit of 1.0. The mean variances are higher than expected; we expect 0.9814, 0.8995 and 0.7002, but we get 0.9867, 0.9349, and 0.8128. These higher values are obtained from both LU and SGS; therefore, we expect that Equation 7 is incorrect. The fitted Gamma distributions appear quite reasonable. A better fit could likely be obtained with a Gaussian distribution when the fluctuations are small. The γ parameter is not that stable; equally good fits could be obtained with different values. The μ parameter decreases with increasing range and the β parameter increases with increasing range, as expected. The fitted parameters for the distribution of variances are shown below:

Case	γ parameter	μ parameter	β parameter
10/50	3.784	0.856	0.035
25/50	3.592	0.673	0.073
50/50	3.650	0.495	0.087

The last row in Figure 1 shows the scatterplot of the realization means versus the realization variance values; note that there is no correlation. The bivariate distribution of the mean and variance could reliably be written as the product of a univariate Gaussian pdf for the mean and a univariate Gamma pdf for the variance. In practice, we would generate a number of realizations to understand the ergodic fluctuations due to the spatial correlation structure, the size of the domain and the conditioning data we have available. For the next step, we may want to have fitted distributions for the mean and variance. Now that we understand ergodic fluctuations we must see how they affect histogram reproduction in original *Z* units.

3 BACK TRANSFORMATION TO ORIGINAL Z UNITS

The back transformation of simulated Gaussian *Y*-values back to original *Z*-units has the potential to introduce some interesting problems. The transformation/back transformation of Equations 1 and 2 may be analytically defined when the *Z*-distribution $F_{Z(z)}$ is defined analytically. In general, however,

the distribution $F_Z(z)$ is defined by a lookup table. Figure 2 shows a schematic illustration that is often used to explain how the transformation works. A one to one reversible transformation is setup using the global distribution. Quantiles of local or conditional distributions are transformed using the global transformation. As mentioned above, most histograms of geologic variables are not Gaussian. We often deal with concentrations that are bounded between 0 and 100%. Variables in small concentration are positively skewed and variables in large concentration tend to be negatively skewed. The affect of a skewed Z distribution on the transformation is the subject of this section.

A simulated realization with a perfectly standard Gaussian distribution (mean of 0 and variance of 1) would be back transformed to the global Z-distribution. Any deviations from a Gaussian shape, a mean of 0 or a variance of 1 would lead to deviations from the global Z distribution. We will be primarily concerned with deviations in the mean and variance (as in the bottom scatterplots on Fig. 1). Figure 3 illustrates the concern of this section. The fluctuations in the mean and variance are understood in Gaussian units (the right side), but we would like to verify that they do not cause any bias in original Z units. Theoretically, we expect no bias, but experience often shows that the back transformed values become biased when care is not taken in Gaussian simulation.

3.1 *Small Example*

We consider positively skewed distributions typified by the lognormal distribution. The lognormal distribution is special because there are analytical links between the Z and Y units. We do not build on those links in this paper because there is no need to limit our results to the lognormal case. For illustration, consider an original Z distribution as lognormal with a mean of 1.0 and a variance of 4.0; the coefficient of variation is 2.0, which is considered typical of a fairly highly skewed distribution. The bivariate distribution of the mean and variance is defined in the previous section. The distribution for the mean is normal and the distribution for the variance is a gamma distribution. The bivariate distribution is the product of these two marginal distributions since we have observed independence between the mean and variance. The bivariate space of the mean and variance is sampled by regular intervals in the Gaussian mean and the Gaussian variance, the values are back transformed and the bivariate probability values are used as weights. The mean/variance relations on the bottom of Figure 1 have been back transformed using this approach (as the schematic on Fig. 3 indicates) using a highly skewed lognormal distribution. Histograms of the mean values are shown on Figure 4. We see no bias in the first case (20% range) and then slight biases in the latter two cases. The magnitude of the bias is not considered significant given the overall spread of the mean values.

The uncertainty in the overall average shown on Figure 4 is very significant, even for the case where the range is only 20% of the field size; although the global average is 1.0 the average can range from 0.5 to 1.5 in the very congenial case where the range is small. There can be no doubt that ergodic fluctuations constitute an important part of uncertainty. It would be an error to transform them away by insisting that each realization reproduce the target mean and variance exactly. Our concern is an overall bias. We conclude that we would see no bias if the mean and variance in Gaussian space were correct. We have seen more significant biases in practical applications. There are a number of practical reasons for biases being introduced into the simulated values in original units.

3.2 *Bias due to Inflated Variance*

The multiple back transformation scheme proposed above can be used to assess the bias due to an inflated variance. The distribution for the mean is left alone and the distribution of variances is increased slightly. We would expect no bias in the resulting mean of the Z values if the original Z distribution is symmetric; we would expect an increasing bias if the original Z distribution is skewed. We can numerically assess the bias given (1) the sampling distributions for the mean and variance, (2) the transformation lookup table, and (3) the increase in the variance. For the sampling distributions from the 50% range case (Fig. 1), a lognormal distribution with a mean of 1 and a variance of 4, and for an increase in variance of 5% we see an increase in the mean of nearly 5% (4.6%). The increase in bias for the lognormal case increases linearly with an increase in variance. We can use this general result to understand the affect of variance inflation due to poor implementation decisions.

4 SOURCES OF BIAS

Most inappropriate implementation choices lead to inflated variance in Gaussian units, which in turn leads to a bias in the mean. We consider the effects of the conditioning data, a small search/small number of data, ordinary kriging, and collocated cokriging. Three courses of action should be considered: (1) careful choice of parameters up front, e.g., choose simple kriging with a large search, (2) careful checking of the results before and after back transformation, and (3) a variance correction within the SGS code, e.g., $\hat{\sigma}^2 = f\sigma^2$ where $f<1$ is applied to every kriging variance at every location used for Gaussian simulation – the value of f must be determined by trial and error using the specific implementation choices.

Conditioning data are essential. They are used to infer the global statistical parameters we need and to constrain the local distributions of uncertainty. The resulting simulated realizations, however, are affected by both the random function model and the conditioning data. The ergodic fluctuations described above in Section 3 apply to unconditional realizations. The fluctuations should decrease with additional conditioning data. The precise decrease in uncertainty could be assessed, but there is no general conclusion. One source of bias due to conditioning data is stationarity. Despite the fact that stationarity is a choice by the practitioner that the data belong to a homogeneous statistical population, large scale trends and border effects (that we call non-stationary features) often cause a bias in the mean and variance. This is hard to illustrate except through anecdotes; however, the results of any study should be checked carefully.

A too-small search and too few data in sequential simulation can result in either an increased or decreased variance that

causes under or over estimation of the global mean when combined with a highly skewed original distribution. Bias in the back transformed histogram is one problem. Another important problem is the poor reproduction of the variogram and volume-variance characteristics of the simulated realizations. In the lognormal scenario developed in Section 3, a reduction in the search radius to ½ of the range or a reduction in the number of data used in kriging leads to a 6.6% bias in the variance.

Ordinary kriging (OK) will certainly lead to an increased variance because of the higher kriging variances that result from implicit estimation of the mean. We have implemented a two-pass simulation procedure where OK is performed for the mean and simple kriging is performed for the variance calculation; it did not work very well. OK may be preferred because it is considered a more robust estimator; however, the global distribution is always used in simulation through the variance and the implicit reliance on the back transformation. In the lognormal scenario developed in Section 3, OK leads to a 5.5% bias in the variance and, consequently, the mean.

Collocated cokriging is particularly troublesome in sequential simulation. The bias in the variance can easily be 20 to 40%, which is not so bad because it is immediately obvious. More subtle biases in the 5 to 10% range are common and not always recognized by the practitioner. Variance correction by an "*f*" factor is almost always required. Biases are particularly bad if the secondary data has greater continuity than the primary variable being simulated and if the absolute value of the correlation is within the 0.5 to 0.9 range.

5 CONCLUSIONS

Geologists and Engineers are being asked to quantify risk and uncertainty. Geostatistical simulation is a powerful tool for that purpose. Gaussian simulation algorithms are the simplest and easiest to apply, yet there are many important implementation details. We make the following conclusions/recommendations:

- Statistical/ergodic fluctuations are a very important factor in uncertainty assessment and they should not be transformed away by insisting that every realization match the global distribution.
- There is no significant bias introduced by the back transform of highly skewed data provided that the ergodic fluctuations are within those expected by the multivariate Gaussian model. Unconditional simulation can be used to understand the reasonably expected fluctuations.
- Forget whatever you learned about kriging. Do not use ordinary kriging; simple kriging is required for unbiased results. Do not limit the search neighborhood; set the neighborhood size equal to the variogram range. Do not limit the number of samples; use at least 24 samples in sequential Gaussian simulation.
- Ergodic fluctuations are not the only source of uncertainty; some form of spatial bootstrap should be considered to assess uncertainty in the target input statistics.

A systematic positive bias in a global resource estimate would be a serious error. Biases can be avoided by careful attention to these implementation decisions and careful checking of the results.

REFERENCES

1. Deutsch, C.V., and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, Second Edition, Oxford University Press, New York, NY, 1998.
2. http://www.itl.nist.gov/div898/handbook/eda/section3/eda366b.htm: NIST Engineering and Statistics Handbook (accessed October 8, 2004).
3. Johnson, N., Kotz, S., and Balakrishnan, N., *Continuous Univariate Distributions, Volumes I and II*, 2nd. Ed., John Wiley and Sons, 1994.
4. Johnson, R., and Wichern, D., *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey, 1998.
5. Journel, A.G., and Huijbregts, Ch., *Mining Geostatistics*, Academic Press, New York, 1978.
6. Journel, A.G., and Xu, W., Posterior Identification of Histograms Conditional to Local Data, *Mathematical Geology*, vol. 26, no. 3, 1994, pp. 323-359.
7. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery B.P., *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing*, Volume 2 of Fortran Numerical Recipes, Cambridge University Press, New York, 2002.

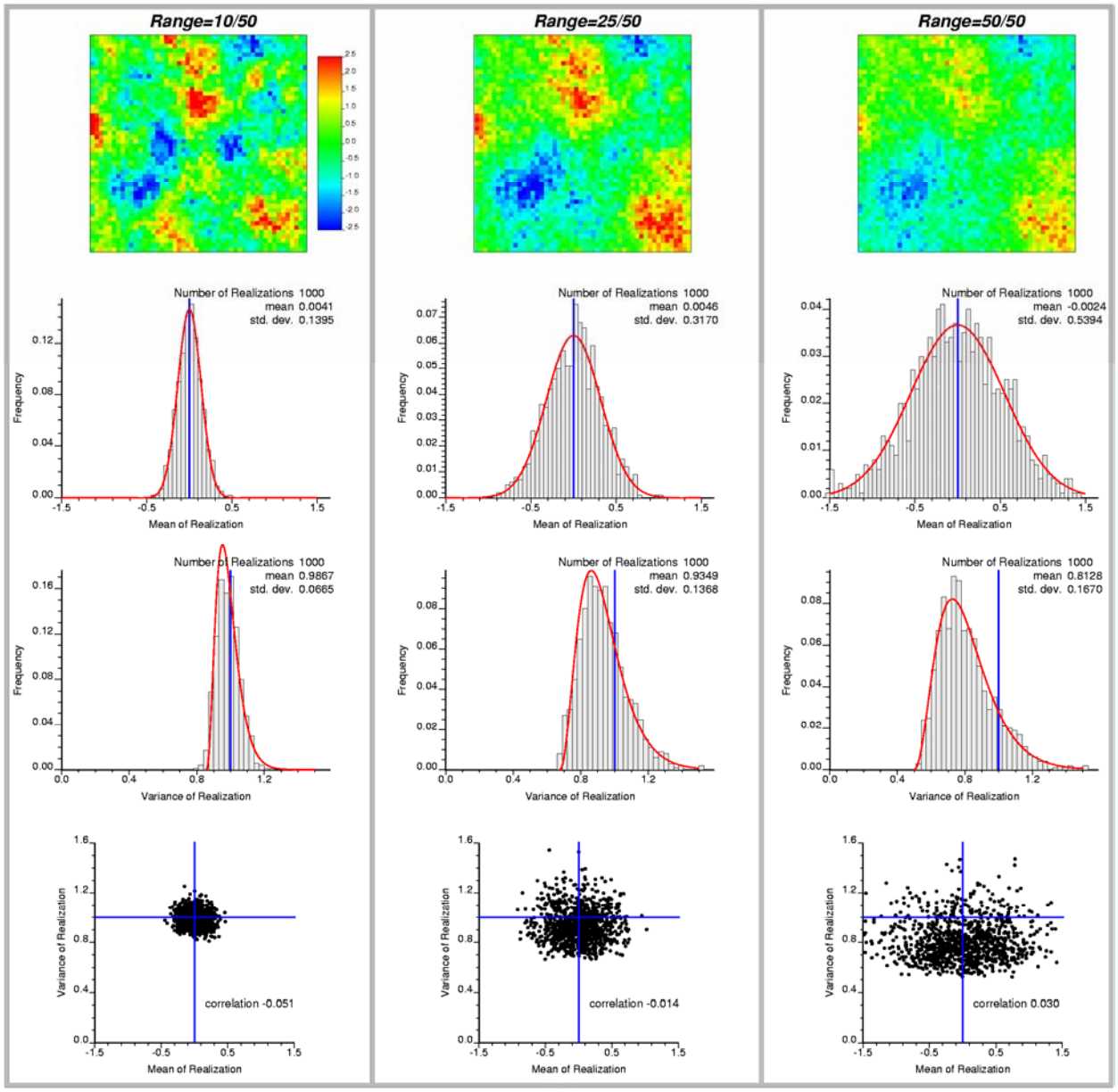

Figure 1: Illustration of ergodic fluctuations for three small 2-D examples with ranges of 20%, 50%, and 100% of the domain size (left to right). The top row shows color scale maps of the first realization, the second row shows the distributions of 1000 mean values, the third row shows the distributions of the 1000 variance values, and the bottom row shows scatterplots of the mean and variance values.

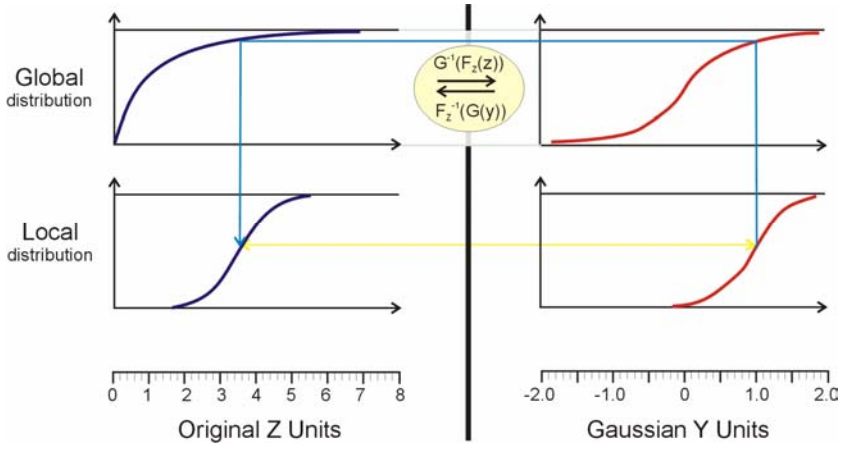

Figure 2: Schematic illustration of normal score transform. The original Z- data are on the left and the Gaussian Y-values are on the right. The top figures are the global CDFs and the bottom figures represent local CDFs. Quantiles are transformed using the global distribution (the three part blue line).

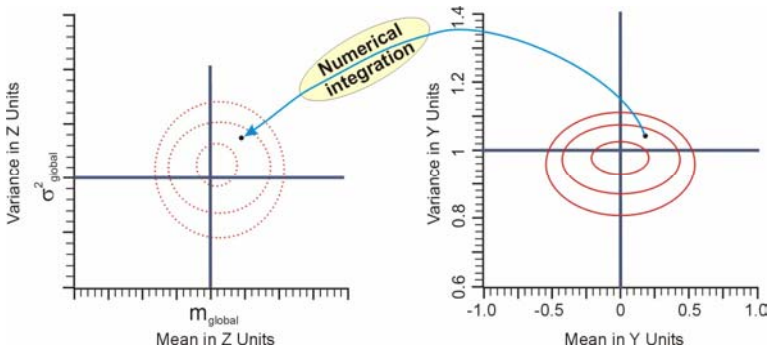

Figure 3: Schematic illustration of how ergodic fluctuations in Gaussian Y units relate back to fluctuations in original Z units. The back transformation must be performed by numerical integration using the transformation illustrated on Figure 2. The back transformation of one realization is illustrated by the light blue arrow with the "Numerical integration" bubble.

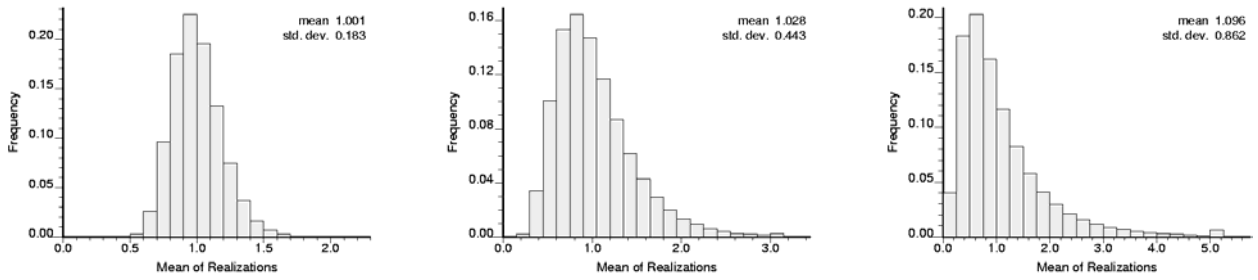

Figure 4: Histograms of the mean values in original units assuming a Z distribution that is lognormal with a mean of 1.0 and a variance of 4.0. The three cases correspond to the ergodic fluctuations on Figure 1, that is, a variogram of 20%, 50% and 100% of the size of the domain. Note the slightly increasing bias and significantly increasing spread as the ergodic fluctuations increase.