Efficient geostatistical simulation for spatial uncertainty propagation

Stelios Liodakis
University of the Aegean
University Hill
Mytilene, Greece
stelioslio@geo.aegean.gr

Phaedon Kyriakidis
Cyprus University of Technology
2-8 Saripolou Str., 3036
Lemesos, Cyprus
phaedon.kyriakidis@cut.ac.cy

Petros Gaganis
University of the Aegean
University Hill, 81100
Mytilene, Greece
gaganis@aegean.gr

Abstract

Spatial uncertainty and error propagation have received significant attention in GIScience over the last two decades. Uncertainty propagation via Monte Carlo simulation using simple random sampling constitutes the method of choice in GIScience and environmental modeling in general. In this spatial context, geostatistical simulation is often employed for generating simulated realizations of attribute values in space, which are then used as model inputs for uncertainty propagation purposes. In the case of complex models with spatially distributed parameters, however, Monte Carlo simulation could become computationally expensive due to the need for repeated model evaluations; e.g., models involving detailed analytical solutions over large (often three dimensional) discretization grids.

A novel simulation method is proposed for generating representative spatial attribute realizations from a geostatistical model, in that these realizations span in a better way (being more dissimilar than what is dictated by change) the range of possible values corresponding to that geostatistical model. It is demonstrated via a synthetic case study, that the simulations produced by the proposed method exhibit much smaller sampling variability, hence better reproduce the statistics of the geostatistical model. In terms of uncertainty propagation, such a closer reproduction of target statistics is shown to yield a better reproduction of model output statistics with respect to simple random sampling using the same number of realizations. The need to process fewer (yet more representative) input realizations implies that the proposed method could contribute to the even wider the application of Monte Carlo-based uncertainty propagation in practice.

Keywords: Monte Carlo simulation, random field models, conditional simulation, Latin hypercube sampling, environmental modeling.

1 Introduction

Spatial uncertainty is endemic in geospatial data due to the imperfect means of recording, processing, and representing spatial information (Zhang and Goodchild, 2002; Shi et al., 2016). As geospatial data often serve as inputs in models with spatially distributed parameters, such as physically-based flow simulation models, the propagation of spatial data uncertainty to uncertainty in model predictions is a critical requirement in GIScience and related fields (Heuvelink, 1998; Caers, 2011).

Although analytical and/or quasi-analytical uncertainty propagation methods have been developed in the literature; see, for example, Şalap-Ayça et al. (2018), Monte Carlo simulation is rather routinely used for uncertainty propagation purposes, as it does not call for, often limiting, assumptions regarding the form of the spatial model itself. In a nutshell, Monte Carlo simulation consists of generating alternative samples (realizations) from the input parameters, evaluating the model response for each of these realizations, and constructing the corresponding distribution of model predictions. The spatial distribution of input parameters is often modeled within a geostatistical framework, and spatial Monte Carlo simulation is performed within the context of geostatistical simulation (Goovaerts, 1997).

Any realistic uncertainty analysis, however, calls for the availability of a representative distribution of model outputs, and can become expensive in terms of both time and computer resources in the case of complex models (Helton and Davis, 2002; Caers, 2011). This problem is far more pronounced in earth and environmental sciences applications, where, in hydrogeology for example, three dimensional grids of hydraulic conductivity values are used along with other parameters to simulate flow and transport in porous media (Gutjahr and Bras, 1993; Chilès and Delfiner, 2012). The computational cost associated with classical Monte Carlo spatial uncertainty propagation calls for the development of more efficient geostatistical simulation methods.

This paper proposes key modifications to classical geostatistical simulation to render it more efficient in terms of generating more representative attribute realizations that better span the range of possible realizations corresponding to a geostatistical specification. Computing model predictions using a set of fewer, yet representative, input parameter realizations, is illustrated to reproduce model output sampling variability corresponding to a much larger input parameter set, thus reducing significantly the computational cost associated with Monte Carlo based spatial uncertainty propagation.

2 Efficient geostatistical simulation

2.1 Geostatistical simulation

In a geostatistical context, the spatial distribution of values of attributes serving as inputs for spatial models is typically conceptualized via a random field; that is, a set of spatially correlated random variables, $\{Y(\boldsymbol{c}), \boldsymbol{c} \in A\}$, one per location (Goovaerts, 1997), where $Y(\boldsymbol{c})$ denotes a random variable (RV) defined at a location with coordinate vector $\boldsymbol{c}$. Geostatistical simulation aims at generating multiple (a set of S) simulated attribute values at a set of M locations $\{\boldsymbol{c}_m, m = 1, \dots, M\}$, typically coinciding with the nodes of a regular grid discretizing the study area; i.e., joint realizations from the M respective RVs $\{Y(\boldsymbol{c}_m), m = 1, \dots, M\}$. Those realizations are often constrained by (or reproduce) N attribute values

$\boldsymbol{y}_1 = [y(\boldsymbol{c}_n), n = 1, \dots, N]^T$ measured at data locations, and the simulation is termed conditional.

In the second-order stationary multivariate Gaussian case (Goovaerts, 1997), the mean of the constituent RVs is assumed constant, $\mu_Y(\boldsymbol{c}_m) = \mu_Y$, and the covariance between any two RVs is a function of the length (distance) and possibly orientation of the vector defined between any two locations; covariance values are typically computed from a distance-decay, positive definite, covariance function inferred from sample data and/or expert knowledge. The *M*-variate joint distribution of the RVs is then Gaussian, and fully characterized by an $(M \times 1)$ constant expectation (mean) vector, denoted as: $\boldsymbol{\mu} = [\mu_Y(\boldsymbol{c}_m), m = 1, \dots, M]^T = \mu_Y \mathbf{1}_M$, where $\mathbf{1}_M$ is a vector with *M* unit entries, and an $(M \times M)$ covariance matrix, $\boldsymbol{\Sigma} = [\sigma_Y(\boldsymbol{c}_m - \boldsymbol{c}_{m'}), m = 1, \dots, M, m' = 1, \dots, M]$, with covariance values between all location pairs.

In the multivariate Gaussian model, the conditional expectation of the *M* RVs given the *N* data values, is furnished by the $(M \times 1)$ vector of Kriging-derived predictions, $\widehat{\boldsymbol{\mu}}_{SK} = [\hat{\mu}_{SK}(\boldsymbol{c}_m), m = 1, \dots, M]^T$, which in the Simple Kriging (SK) case is written as:

$$\widehat{\boldsymbol{\mu}}_{SK} = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}[\boldsymbol{y}_1 - \boldsymbol{\mu}_1]$$

where $\boldsymbol{\mu}_1 = [\mu_Y(\boldsymbol{c}_n), n = 1, \dots, N]^T$ is a $(N \times 1)$ vector of expected values at the data locations, $\boldsymbol{\Sigma}_{11}^{-1}$ is the inverse of the $(N \times N)$ matrix $\boldsymbol{\Sigma}_{11} = [\sigma_Y(\boldsymbol{c}_n - \boldsymbol{c}_{n'}), n = 1, \dots, N, n' = 1, \dots, N]$ of covariance values between all data location pairs, and $\boldsymbol{\Sigma}_{21} = [\sigma_Y(\boldsymbol{c}_m - \boldsymbol{c}_n), m = 1, \dots, M, n = 1, \dots, N]$ is the $(M \times N)$ matrix of covariance values between all simulation and data location pairs.

The uncertainty in the Simple Kriging predictions is encapsulated in the $(M \times M)$ matrix $\widehat{\boldsymbol{\Sigma}}_{SK} = [\sigma_Y(\boldsymbol{c}_m - \boldsymbol{c}_{m'}), m = 1, \dots, M, m' = 1, \dots, M]$ of conditional (co)variance values between all simulation location pairs, given as:

$$\widehat{\boldsymbol{\Sigma}}_{SK} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

where $\boldsymbol{\Sigma}_{12} = [\sigma_Y(\boldsymbol{c}_n - \boldsymbol{c}_m), n = 1, \dots, N, m = 1, \dots, M] = \boldsymbol{\Sigma}_{21}^T$ is the $(N \times M)$ matrix of covariance values between all data and simulation location pairs. Note that the diagonal entries of matrix $\widehat{\boldsymbol{\Sigma}}_{SK}$ are the values of the SK prediction error variance at the *M* simulation locations.

Geostatistical simulation aims at generating (or sampling) a set of *S* attribute realizations (e.g., images in 2D), from the multivariate Gaussian distribution function $G(\widehat{\boldsymbol{\mu}}_{SK}, \widehat{\boldsymbol{\Sigma}}_{SK})$. More specifically, a $(S \times M)$ matrix $\boldsymbol{Y} = [y_s(\boldsymbol{c}_m), s = 1, \dots, S, m = 1, \dots, M]$ with conditionally simulated attribute values (a Simple Random (SR) sample of size *S*) – whose *s*-th row contains one conditional realization, i.e., one simulated attribute value for each of the *M* locations, and whose *m*-th column contains *S* simulated attribute values at one location -- can be generated as (Goovaerts, 1997):

$$\boldsymbol{Y} = \widehat{\boldsymbol{M}}_{SK} - \boldsymbol{W}\widehat{\boldsymbol{L}}_{SK}^T$$

where $\widehat{\boldsymbol{M}}_{SK}$ is a $(S \times M)$ matrix with *S* replicates of vector $\widehat{\boldsymbol{\mu}}_{SK}^T$ along its *S* rows, and $\boldsymbol{W}\widehat{\boldsymbol{L}}_{SK}^T$ is a $(S \times M)$ matrix of SK prediction error realizations (one per row), with $\boldsymbol{W} = [w_s(\boldsymbol{c}_m), s = 1, \dots, S, m = 1, \dots, M]$ being a $(S \times M)$ matrix of standard Gaussian deviates and $\widehat{\boldsymbol{L}}_{SK}$ being the $(M \times M)$ lower triangular Cholesky factor of the conditional covariance matrix $\widehat{\boldsymbol{\Sigma}}_{SK}$.

2.2 Latin hypercube sampling for efficient geostatistical simulation

An efficient alternative to classical Monte Carlo simulation based on SR sampling is Latin hypercube (LH) sampling, a form of stratified random sampling, aiming at generating representative samples or realizations from a set of random variables with a given multivariate probability distribution (McKay, Beckman and Conover, 1979). LH sampling has been shown to lead to model outputs with smaller sampling variability in their statistics than SR sampling for the same number of input simulated realizations; that efficiency, however, decreases the more non-linear that model becomes in the parameters (Helton and Davis, 2003).

The most widely used methods for generating LH samples from a multivariate distribution are those of Iman and Conover (1982) and Stein (1987). In the first method, the entries of an uncorrelated SR or LH sample are re-arranged to match a target rank correlation matrix. In the second method, a correlated SR sample is transformed into a correlated LH sample based on the ranks of the former; correlation is inherited in the LH sample from the correlation in the ranks of the original SR sample. These methods do not rely on any Gaussian assumption, and both can be used for simulation with or without conditioning data. Relevant representative applications in a spatial context include the work of Zhang and Pinder (2003) and Pebesma and Heuvelink (1999), respectively.

In what follows, classical, SR-sampling based, geostatistical simulation is modified to incorporate Stein's LH sampling algorithm. More precisely, the local SR sample comprised of *S* conditionally simulated attribute values at any location $\boldsymbol{c}_m$ is transformed into a conditional LH sample, comprised of *S* values stratified into *S* strata, one per stratum, as:

$$y_s^*(\boldsymbol{c}_m) = G^{-1}\left(\frac{r(y_s(\boldsymbol{c}_m)) - v_s(\boldsymbol{c}_m)}{S}; \hat{\mu}_{SK}(\boldsymbol{c}_m), \hat{\sigma}_{SK}(\boldsymbol{c}_m)\right)$$

where G^{-1} is the inverse conditional cumulative distribution function (here Gaussian) of RV $Y(\boldsymbol{c}_m)$ with parameters $\hat{\mu}_{SK}(\boldsymbol{c}_m)$, the local Kriging prediction, and $\hat{\sigma}_{SK}(\boldsymbol{c}_m)$, the local Kriging prediction error variance, $r(y_s(\boldsymbol{c}_m))$ denotes the rank of the simulated attribute value $y_s(\boldsymbol{c}_m)$ ranging from *1* for the smallest to *S* for the largest value simulated at the same location, and $v_s(\boldsymbol{c}_m)$ is a random number uniformly distributed in the $[0, 1]$ interval.

The rank value $r(y_s(\boldsymbol{c}_m))$ identifies a probability stratum associated with an original simulated value $y_s(\boldsymbol{c}_m)$, and $v_s(\boldsymbol{c}_m)$ furnishes a random probability perturbation within that stratum. The stratified probability values are then transformed into stratified Gaussian quantiles via the inverse local CDF. The result is a conditionally simulated LH sample of size *S*, marginally (location-wise) stratified, thus avoiding too similar by chance simulated values. Spatial correlation is induced in the LH realizations via the

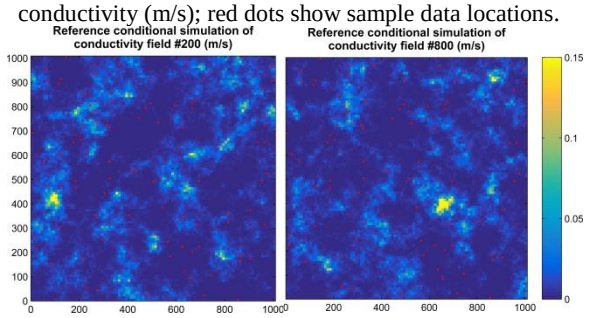
ranks of the original (generated via SR sampling) conditionally simulated attribute values.

3 Synthetic case study

The application of the proposed conditional Latin hypercube simulation method, in comparison with SR sampling, is illustrated via a synthetic case study within a hydrogeological context involving flow and transport in a heterogeneous porous medium. More precisely, a two-dimensional synthetic groundwater flow system is considered, similar to that adopted by Zhang and Pinder (2003) and Kyriakidis and Gaganis (2013). The dimensions of the flow system are *1010m* by *1010m* discretized into a *101×101* grid with uniform rectangular cells of size *10m* by *10m* each.

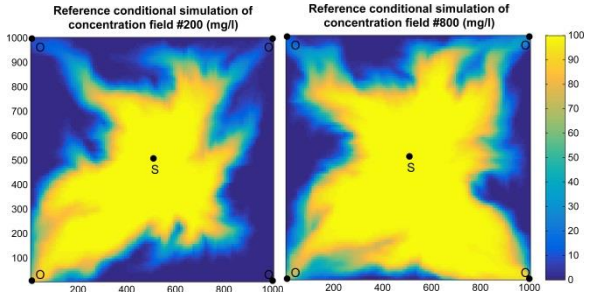
The spatial distribution of hydraulic conductivity values in this domain is modeled as a realization of a second-order stationary and isotropic lognormal random field with parameters (mean and variance) derived from real-world data reported in Sudicky et al. (2010). The semivariogram of log conductivity is assumed to be of exponential form, with no nugget effect, and effective range 202 m, corresponding to one fifth of the domain extent along the cardinal directions. Sample hydraulic conductivity data were extracted from an unconditional simulation from that reference random field model, and were used as conditioning data for all subsequent simulations. Two (reference) conditional realizations of this random field model using geostatistical simulation with SR sampling are illustrated in Figure 1.

Figure 1: Two conditional simulations of hydraulic conductivity (m/s); red dots show sample data locations.



A set of *1000* hydraulic conductivity (reference) conditional simulations were generated using SR sampling and used as input for a flow and transport model. More precisely, flow boundary conditions consisted of constant hydraulic head *0m* at the four corner cells and a constant hydraulic head of *50m* at the central cell of the domain; no flow conditions were assigned to the rest of the domain boundaries. For the solute transport problem, an initial concentration equal to *0mg/l* is assumed throughout the model domain. At time $t = 0$, a contaminant is introduced at the central source cell, along the upstream constant head boundary, with constant concentration *100mg/l*; no transport conditions are assigned along the domain boundaries. In terms of software, the Modflow code (McDonald and Harbaugh, 1988) was used to obtain the steady state flow solution, and the MT3D code (Zheng, 1990) was used to obtain the solute transport solution up to time $t =$ *2·10^6sec*. Two simulated solute concentration realizations, the ones corresponding to the hydraulic conductivity conditional realizations of Figure 1, are shown in Figure 2.

Figure 2: Solute concentration realizations (mg/l); S denotes the contaminant source location.



Ensemble statistics for hydraulic conductivity and solute concentration are derived from the set of *1000* conditional realizations of conductivity (representing model inputs) and the corresponding set of *1000* simulations of concentration (representing model outputs), respectively. The ensemble average and standard deviation fields for conductivity pertain to the mean and standard deviation of simulated conductivity values at any location, and are show in Figure 3. Similarly, the ensemble mean and standard deviation fields for concentration pertain to simulated concentration values at any location, and are show in Figure 4. These four ensemble fields (two for conductivity and two for concentration) are considered as reference statistics, as they are derived from a very large set of simulations.

Figure 3: Ensemble average (left) and standard deviation (right) fields of hydraulic conductivity (m/s).

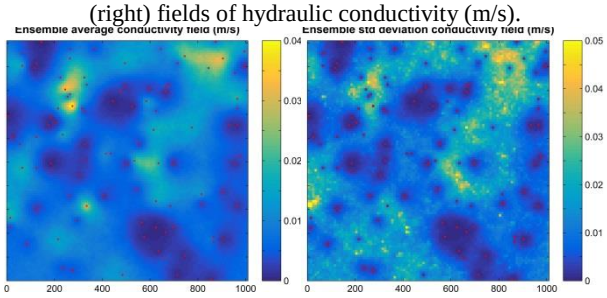
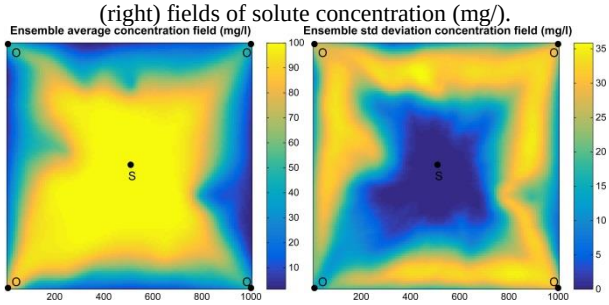


Figure 4: Ensemble average (left) and standard deviation (right) fields of solute concentration (mg/).



The objective now becomes the reproduction of the ensemble statistics from analogous statistics computed from sets of simulations involving much fewer realizations. More

precisely, conditional simulation using classical SR sampling and the proposed LH sampling are considered for generating realizations of a lognormal hydraulic conductivity field, with the same parameters used for the reference simulations, and two sample sizes (number of realizations); namely, $S = 10$ and $S = 30$. Once a sample is generated, the discrepancy between the statistics of the simulated ensemble and the reference ensemble statistics fields is quantified using the root mean squared error (RMSE). The computation of RMSE is repeated over a set of *100* batches of realizations, with each batch containing the same sample size, in order to compute the sampling distribution of the RMSE. The sampling distributions of RMSE values for each sample size and for each method are then used to compare the reproductive abilities of the methods under consideration.

In Figures 5 and 6 hereafter, RMSE sampling distributions are presented in terms of their means and medians, as well as their *75%* and *95%* probability intervals. Mean values are depicted with circles (o), median values with asterisks (*), 75% RMSE probability intervals with horizontal line segments, and 95% probability intervals with × symbols. The better the reproduction of a reference ensemble statistic field from realizations of a simulation method, for a given sample size, the narrower the sampling distribution of the resulting RMSE values, the smaller (closer to 0) the center of that distribution.

Figure 5: Hydraulic conductivity ensemble average (left) and standard deviation (right) reproduction.

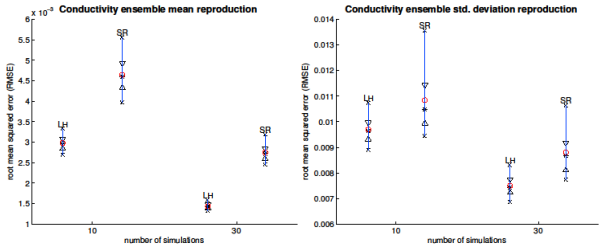
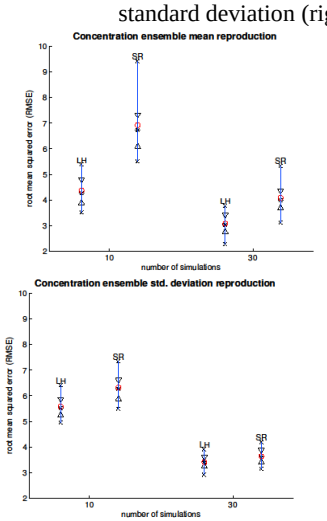


Figure 6: Solute concentration ensemble average (left) and standard deviation (right) reproduction



It can be readily appreciated that the proposed LH-based geostatistical simulations of hydraulic conductivity reproduce much better the reference ensemble statistics fields of Figure 3 for both sample sizes that the classical SR-based simulations. In other words, the centers of the corresponding vertical bars of Figure 5 are much lower for the proposed method than for classical SR sampling; this implies that the ensemble statistics from the proposed method are more similar to the corresponding reference statistics. Moreover, those bars are also narrower for the proposed method implying a smaller spread; i.e., smaller sampling variability and closer agreement with the reference target statistics. Similar conclusions can be reached for the case of solute concentration from Figure 6, although the differences between the proposed and classical methods are smaller, particularly for the case of the ensemble standard deviation field (Figure 6, right).

4 Discussion

Monte Carlo simulation based on simple random (SR) sampling is typically the method of choice for uncertainty propagation purposes in GIScience and related applications in earth and environmental sciences. This method, however, can become computationally expensive in the case of models with spatially distributed parameters or inputs, such as complex environmental models involving flow and transport, whereby repeated evaluation of computationally expensive models is required.

A novel geostatistical simulation method has been proposed in this paper, whereby the concept of Latin hypercube (LH) sampling, widely used in a non-spatial context, is integrated in classical conditional geostatistical simulation. The result is a computationally efficient method for generating representative (not too similar by chance) realizations of spatial variables from a random field model; these realizations are shown to span in a much better way the "uncertainty space" pertaining to both the input variables and the output model predictions. It is expected that the proposed geostatistical simulation method will contribute to an even wider application of Monte Carlo based spatial uncertainty propagation in practice.

As the proposed geostatistical simulation method in this work involves the Cholesky factorization of a covariance matrix, a task that becomes prohibitive for simulation at a large number (>50,000) of locations, further enhancements are required to tackle the issue of simulation at the nodes of large (often 3D) discretization grids. Such extensions have been reported in Liodakis et al. (2015) for the unconditional simulation case, and are currently under development for the conditional simulation case.

References

Caers, J., (2011) *Modeling Uncertainty in the earth sciences*, New York, Wiley.

Chilès, J.-P., and Delfiner, P. (2012) *Geostatistics: modeling spatial uncertainty*, Second Edition, Hoboken, Wiley.

Goovaerts, (1997) *Geostatistics for natural resources evaluation*, New York, Oxford University Press.

Gutjahr, A.L., and Bras, R.L. (1993) Spatial variability in subsurface flow and transport: A review, *Reliability Engineering and System Safety*, 42, 293-316.

Helton, J.C., and Davis, F.J. (2002) Illustration of sampling-based methods for uncertainty and sensitivity analysis, *Risk Analysis*, 22(3), 591-622
Helton, J.C., and Davis, F.J. (2003) Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, *Reliability Engineering & System Safety*, 81(1), 23-69.

Heuvelink, G. (1998) *Error Propagation in Environmental Modeling with GIS*, Taylor & Francis, London.

Iman, R.L., and Conover, W.J. (1982) A distribution-free approach to inducing rank correlation among input variables, *Communications in Statistics, Part B–Simulation and Computation*, 11(3), 311-334.

Kyriakidis, P., and Gaganis, P. (2013) Efficient simulation of (log)normal random fields for hydrogeological applications, *Mathematical Geosciences*, 45(5), 531-556.

Liodakis, S., Kyriakidis, P., and Gaganis, P. (2015) Effcient uncertainty propagation of lognormal hydraulic conductivity in a three dimensional hydrogeological model of flow and transport on very large regular grids. In: *Proceedings of the 17th Annual Conference of the International Association for Mathematical Geosciences (IAMG 2015)*, Freiberg, 2015.

McDonald, M., and Harbaugh, A. (1988) A modular three-dimensional finite difference ground-water flow model, Technical Report, *Techniques of Water-Resources Investigations, Book 6: Modeling Techniques*, U.S. Geological Survey.

McKay, M.D., Beckman, R.J., and Conover, W.J. (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, 21(2), 239-245.

Pebesma, E.J., and Heuvelink, G.B.M. (1999) Latin hypercube sampling of Gaussian random fields, *Technometrics*, 41(4), 303-312.

Şalap-Ayça, S., Jankowski, P., Clarke, K.C., Kyriakidis, P.C., and Nara, A. (2018) A meta-modeling approach for spatio-temporal uncertainty and sensitivity analysis: an application for a cellular automata-based urban growth and land-use change model, *International Journal of Geographical Information Science*, 32(4), 637-662.

Shi, W., Wu, B., and Stein, A. (2016) *Uncertainty modelling and quality control for spatial data*, Boca Raton, CRC Press.

Stein, M. (1987) Large sample properties of simulations using Latin hypercube sampling, *Technometrics*, 29(2), 143-151.

Sudicky E., Illman, W., Goltz, I., Adams, J., McLaren, R. (2010) Heterogeneity in hydraulic conductivity and its role on the macroscale transport of a solute plume: From measurements to a practical application of stochastic flow and transport theory, *Water Resources Research*, 46(1), W01508, DOI: 10.1029/2008WR007.

Zhang, J., and Goodchild, M.F. (2002) *Uncertainty in geographical information*, London, Taylor & Francis.

Zhang, Y., and Pinder, G. (2003) Latin hypercube lattice sample selection strategy for correlated random hydraulic conductivity fields, *Water Resources Research*, 39(8), DOI: 10.1029/2002WR001822.