Geostatistical Sampling Methods for Efficient Uncertainty Analysis in Flow and Transport Problems

Stelios C. Liodakis[*a], Phaedon C. Kyriakidis[a,b], Petros Gaganis[c]*

[a] Department of Geography, University of the Aegean (University Hill, Mytilene, Lesvos, 81100, Greece; E-mail: stelioslio@geo.aegean.gr, phkyriakidis@geo.aegean.gr)
[b] Department of Geography, University of California, Santa Barbara (CA 93106-4060, USA; E-mail: phaedon@geog.ucsb.edu)
[c] Department of Environmental Studies, University of the Aegean (University Hill, Mytilene, Lesvos, 81100, Greece; E-mail: gaganis@aegean.gr)

This poster participates in OSP

Outstanding Student Poster Contest

1. Introduction

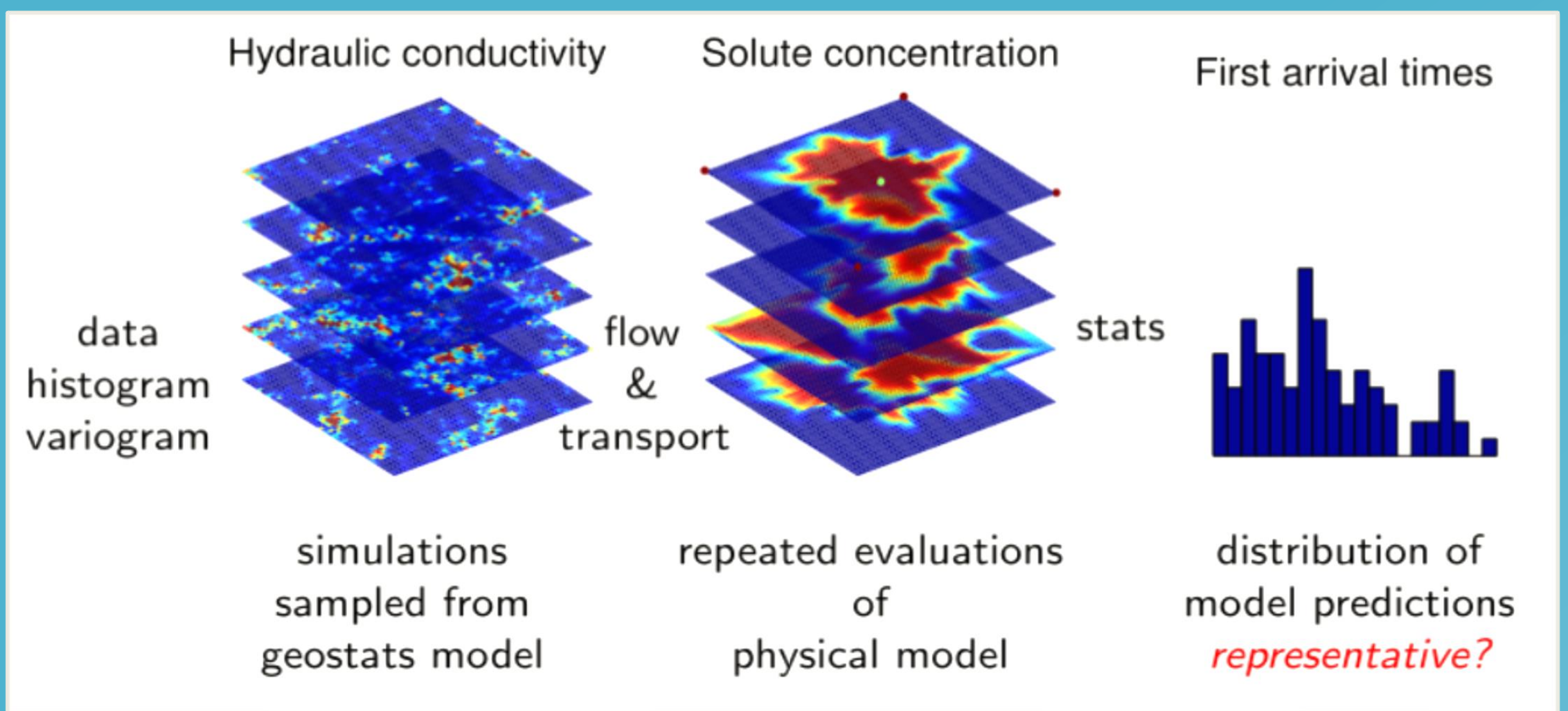

Figure 1. Workflow for Monte Carlo uncertainty propagation

Uncertainty analysis in hydrogeological investigations involving flow and transport in heterogeneous porous media is often conducted in a Monte Carlo framework to evaluate, for example, the uncertainty in the spatial distribution of solute concentration due to the uncertainty in the spatial distribution of hydraulic conductivity. In this context, the spatial distribution of hydraulic conductivity is frequently parameterized in terms of a lognormal random field model, from which simulated conductivity realizations are often generated via geostatistical simulation involving simple random (SR) sampling from the multivariate (log)normal distribution [1]. Realistic uncertainty analysis, however, calls for a large number of simulated conductivity fields; hence, can become expensive in terms of both time and computer resources.

A more efficient alternative to SR sampling is Latin hypercube (LH) sampling, a special case of stratified random sampling, which yields a more representative distribution of simulated attribute values with fewer realizations [2]. Here, term representative implies realizations spanning efficiently the range of possible attribute values corresponding to the multivariate (log)normal probability distribution associated with the random field model.

2. Materials and Methods

In this work we investigate the efficiency of alternative methods to classical LH sampling within the context of simulation of flow and transport in a heterogeneous porous medium. More precisely, we consider the stratified likelihood (SL) sampling method of [3], in which attribute realizations are generated using the polar simulation method by exploring the geometrical properties of the multivariate Gaussian distribution function. In addition, we propose a more efficient version of the above method, here termed minimum energy (ME) sampling. In both cases a set of N representative conductivity realizations at M locations is constructed by: (i) generating a representative set of N points distributed on the surface of a M-dimensional, unit radius hyper-sphere, (ii) relocating the N points on a representative set of N hyper-spheres of different radii, and (iii) transforming the coordinates of those points to lie on N different hyper-ellipsoids spanning the multivariate Gaussian distribution [4] (Figure 2).

The above steps for generating a stratified sample of size N from a multivariate Gaussian PDF are summarized as $Y = diag(x)\tilde{U}C + M_Y$, where $diag(x)$ is a $(N \times N)$ diagonal matrix having as diagonal entries a set of N deviates $x = [x_n, n = 1,\dots,N]^T$ from a chi distribution with K degrees of freedom, term $\tilde{U}$ is a $(N \times K)$ matrix with uniform deviates in $[-1, 1]$, all rows of which are constrained to have a unit norm and M_Y is an expectation matrix $(N \times K)$.

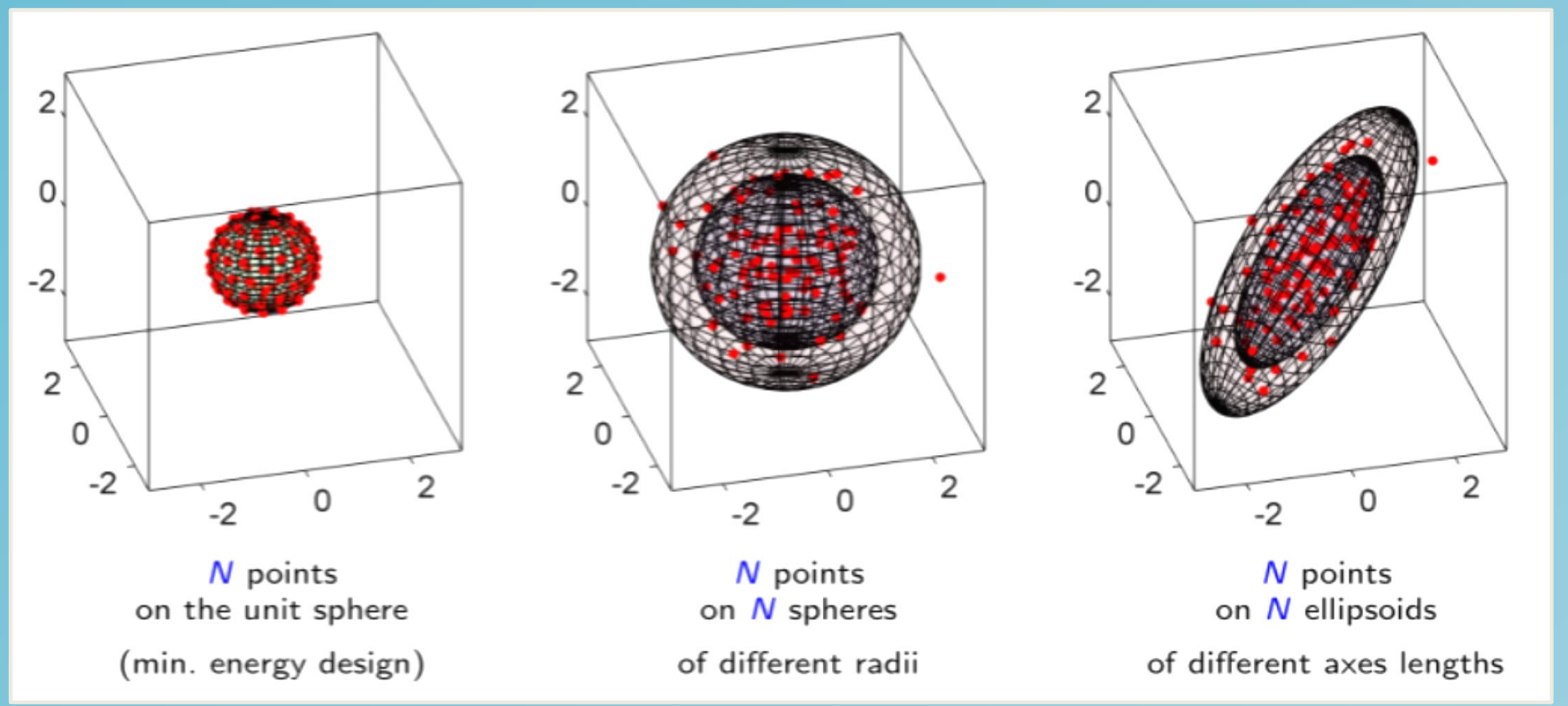

Figure 2. Three dimensional depiction for generating a stratified sample of size N from a multivariate Gaussian PDF

SL sampling creates $\tilde{U}$ by generating a $(N \times K)$ matrix V_L with a stratified sample of size N from K uncorrelated standard Gaussian deviates, computing the vector $\| V_L \|$ with the N scalar norms of the rows of V_L, and computing the $(N \times K)$ matrix $\tilde{U}_L = diag(\| V_L \|)^{-1} V_L$; the N rows of $\tilde{U}$ approximate N stratified, unit norm, realizations or points on the surface of a unit hyper-sphere in a K dimensional hypersphere. SL sampling does not guarantee an optimal (with minimum energy) placing the N points on the surface of a unit (hyper)sphere.

ME sampling first generates N random points on the surface of the unit (hyper)sphere. Then a repulsive force vector, based on $1/r^2$, where r denotes the smallest linear distance between two neighbouring points, is calculated for each point. The resultant force vector is normalized, and then each point is displaced a distance $S = 1$ in the direction of that force, and finally projected back down onto the unit sphere [7]. When the system nears convergence, the displacement vector for a given point is nearly in the same direction as the radius vector for that point due to the points being equally distributed [5]. The steps for generating N stratified points using ME sampling are shown in Figure 3, where the point, depicts the transition from random points (small) to the system minimum energy convergence (large).

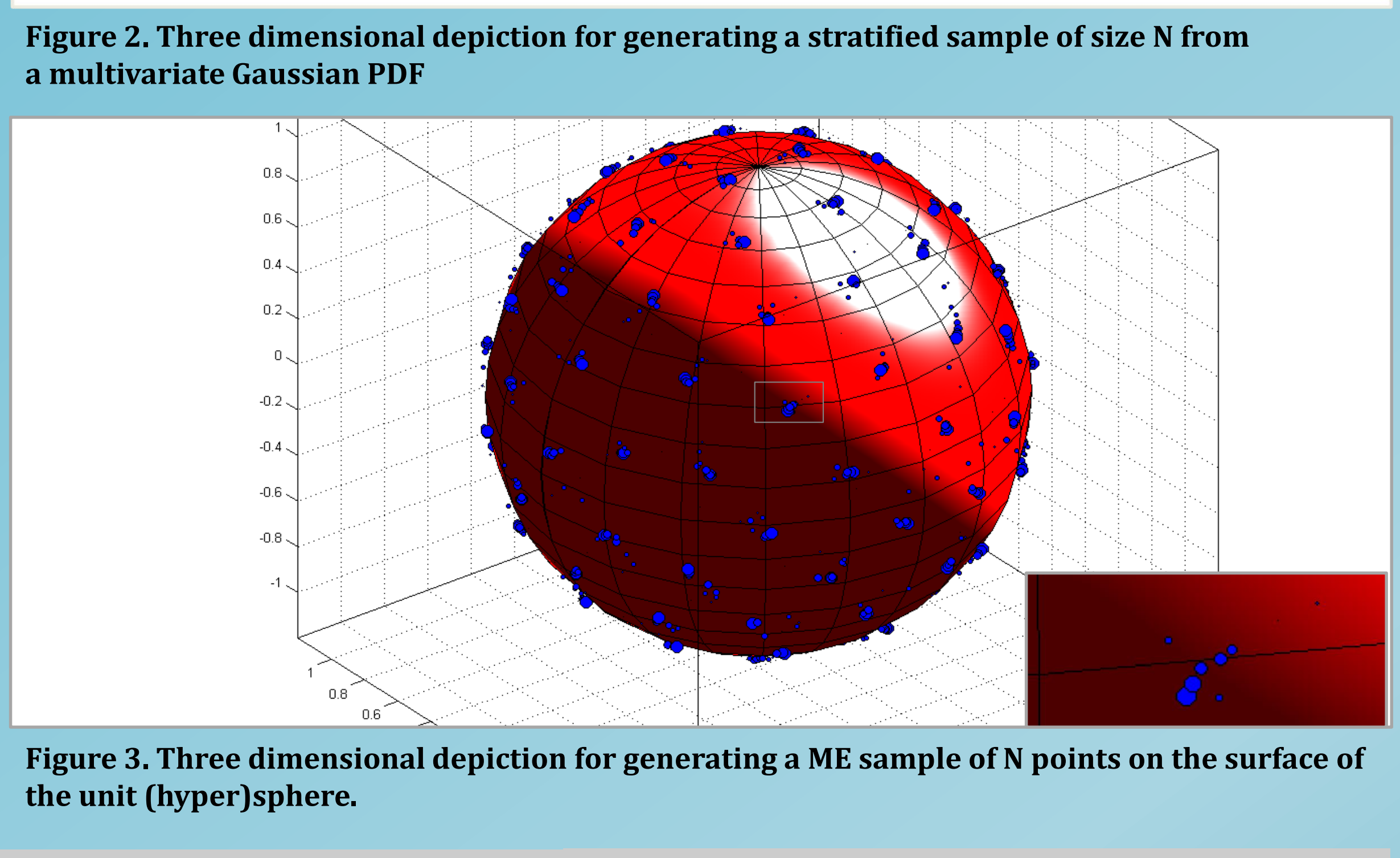

Figure 3. Three dimensional depiction for generating a ME sample of N points on the surface of the unit (hyper)sphere.

The ability of the three stratified sampling methods (LH-SL-ME) considered in this study, at furnishing representative attribute values, in terms of maximum dissimilarity between them, was explored by generating correlated hydraulic conductivity values at nine control points. The results are illustrated in Figure 4A, where for both sample sizes under consideration – *10 / 30* – ME sampling displays the largest nearest neighbour dissimilarities in simulated hydraulic conductivity values. Figure 4B depicts the dissimilarities of concentration values at the same control points, resulting from the hydrogeological model evaluation.

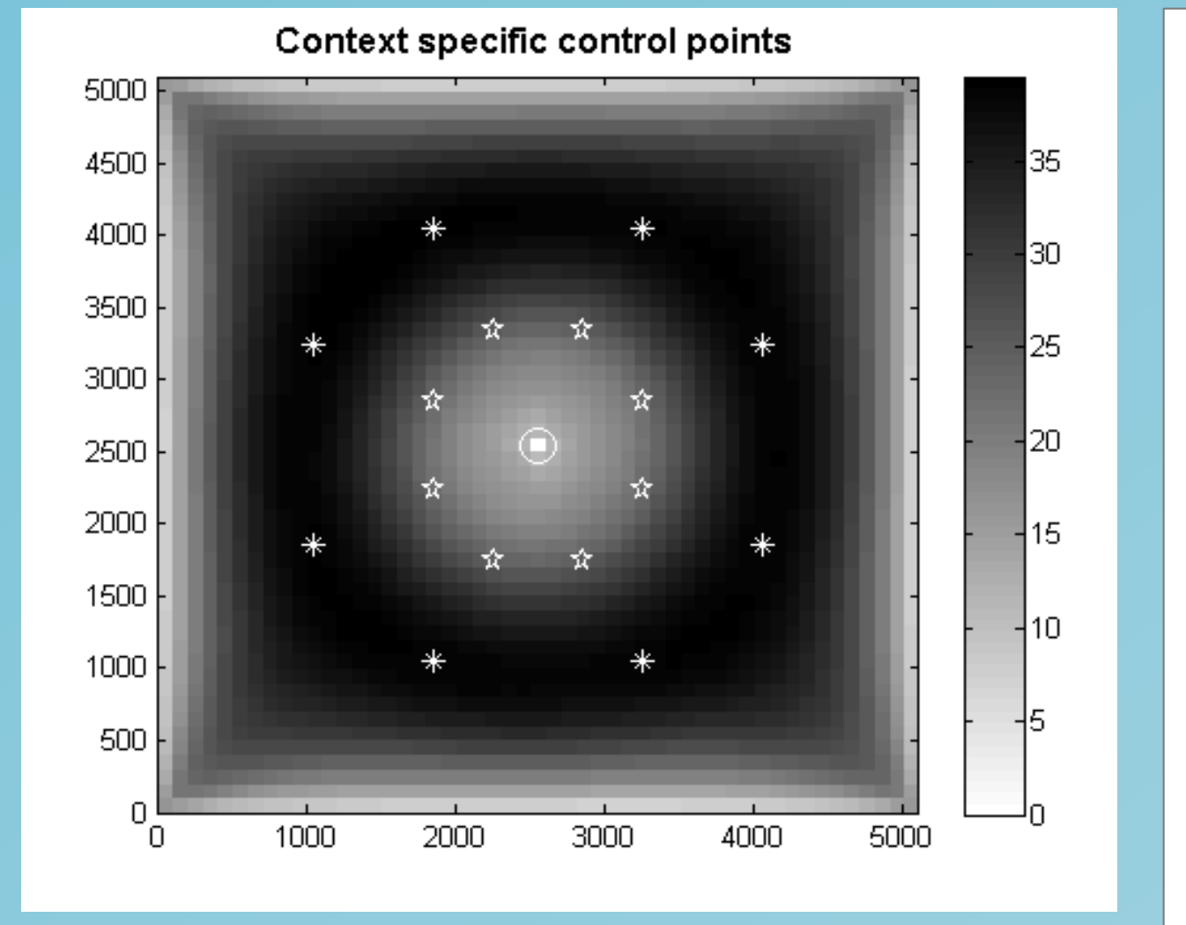

Figure 5. Application - specific control points plotted on the ensemble standard deviation field of concentration

The above methods, along with the LH sampling method are applied in a dimensionality reduction context by selecting flow-controlling points over which representative sampling of hydraulic conductivity is performed, thus also accounting for the sensitivity of the flow and transport model to the input hydraulic conductivity field [6]. According to [4], one could consider control points at regions of highest uncertainty in terms of data control, or alternatively in terms of model response uncertainty. Moreover, control points should correspond to application-specific important locations in terms of controlling the variance of realizations of model outputs; in this case, the ensemble standard deviation of solute concentration (Figure 5).

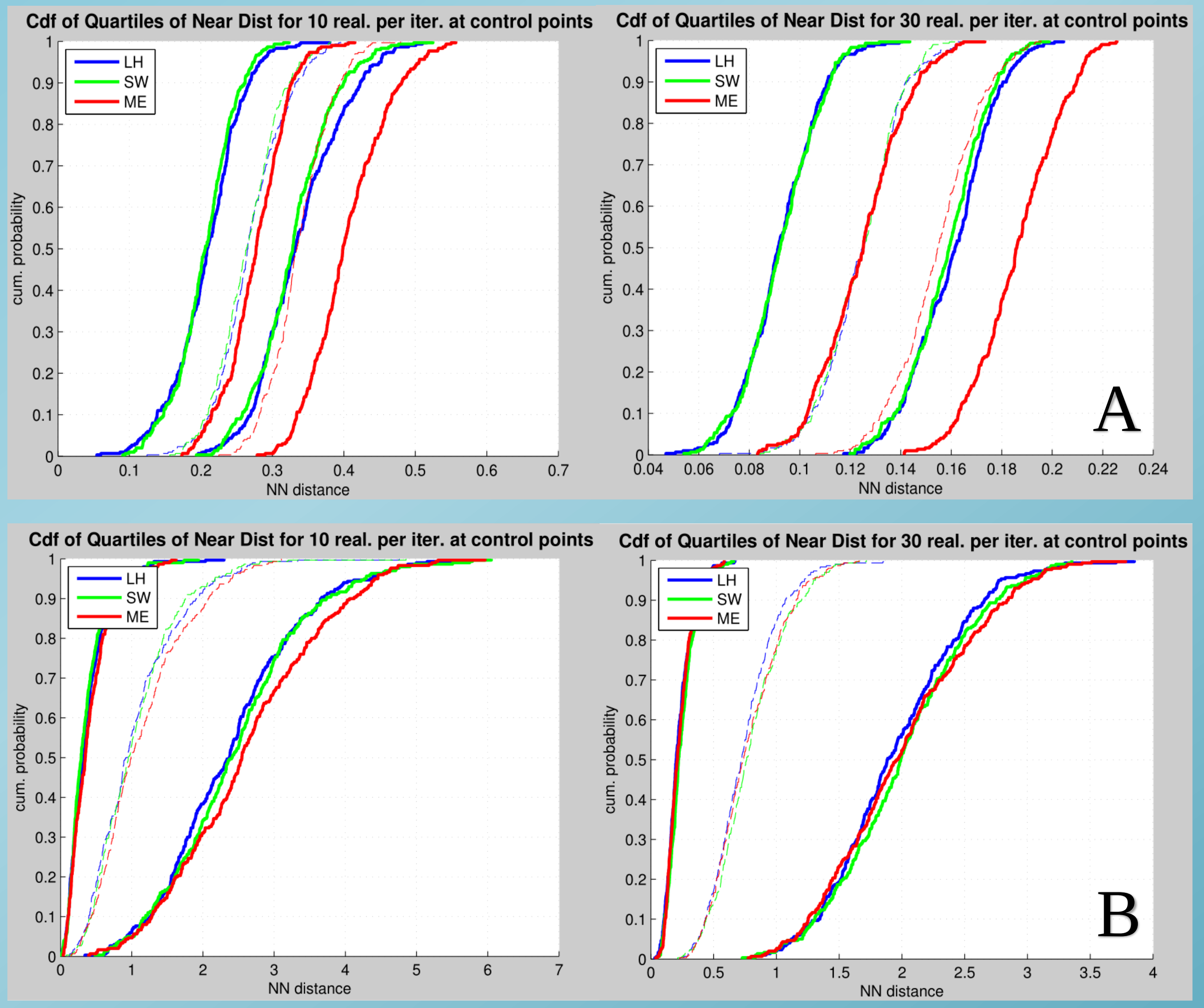

Figure 4. CDF of nearest neighbour distances – dissimilarities between simulated values of A) hydraulic conductivity (cosine metric) and B)concentration (mahalanobis metric), at nine control points (eight points radially emanating along the first areola depicted as stars along the central point depicted as circle at Figure 5)

Acknowledgements

The research project (2269) "Advances in Geostatistics for Environmental Characterization and Natural Resources Management" with acronym "GEOSTATENV" is implemented within the framework of the Action "Aristeia I" of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.

European Union
European Social Fund
OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS
MANAGING AUTHORITY
NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND
Co-financed by Greece and the European Union

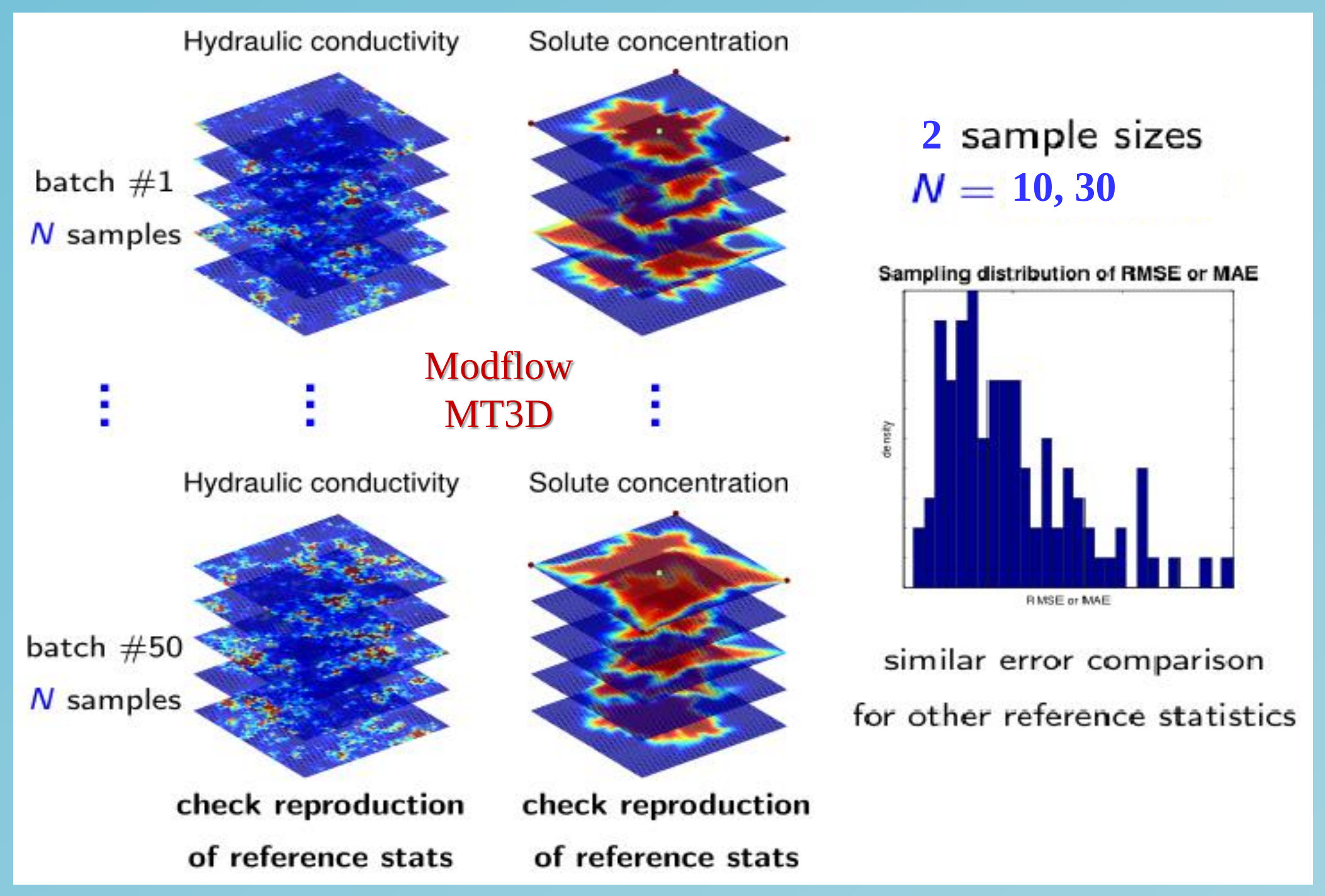

Figure 6. Workflow for evaluating the performance of simulation methods.

The performance of sampling methods, LH, SL, and ME, is compared for different sample sizes N (10 -30), to that of SR sampling in terms of reproduction of ensemble statistics of reference (10000 SR realizations) hydraulic conductivity and solute concentration fields (Figure 7)

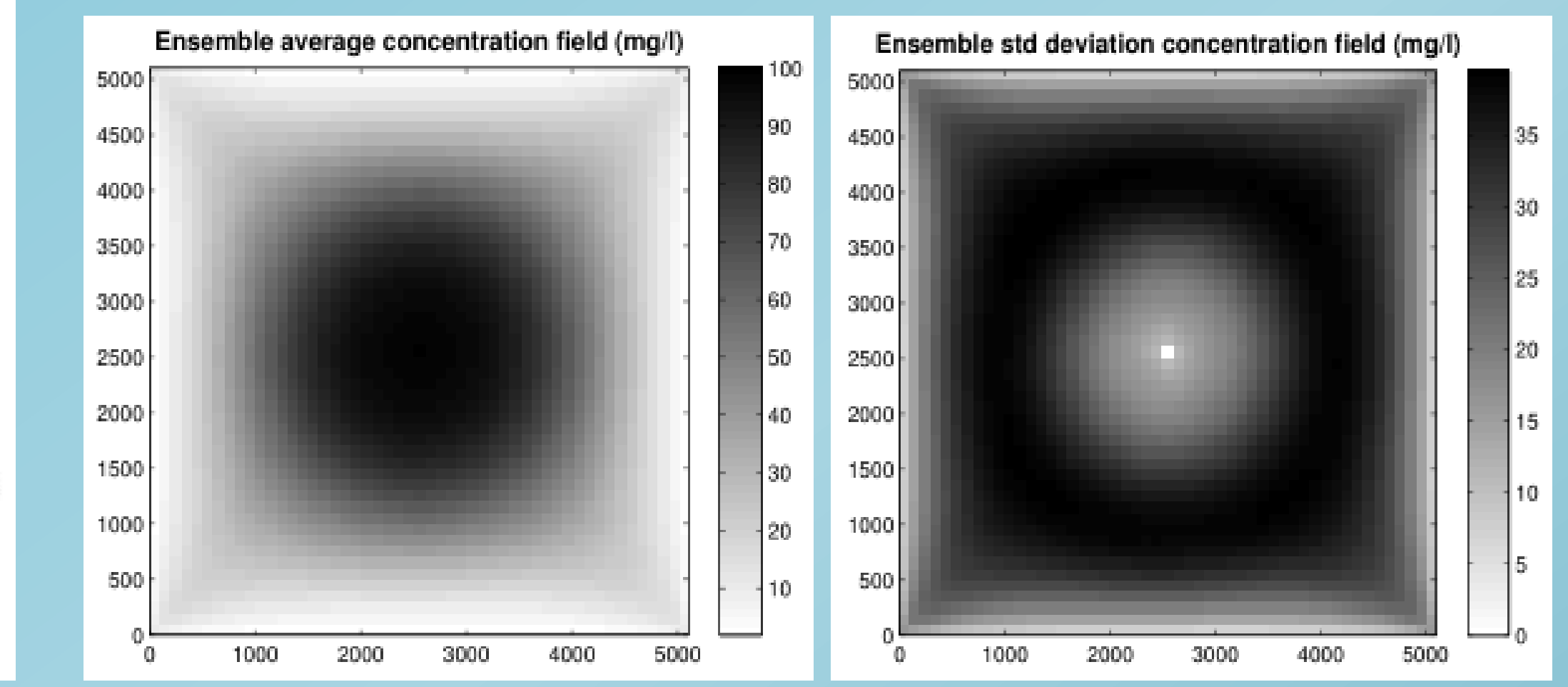

Figure 7. Ensemble mean (left) and standard deviation (right) reference concentration fields .

3. Results

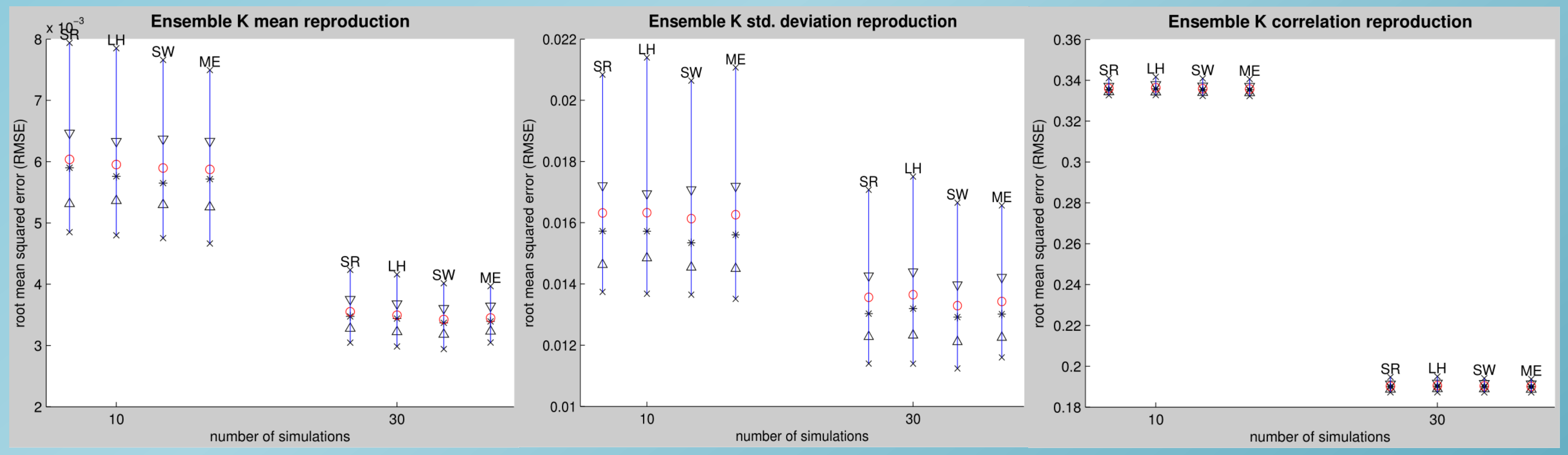

Figure 8. Reproduction of ensemble statistics i) mean, ii) std deviation and iii) correlation, between hydraulic conductivity realizations and the ensemble average hydraulic conductivity field. Reproduction is quantified here in terms of the sampling distribution of RMSE between reference and simulated ensemble concentration statistics .

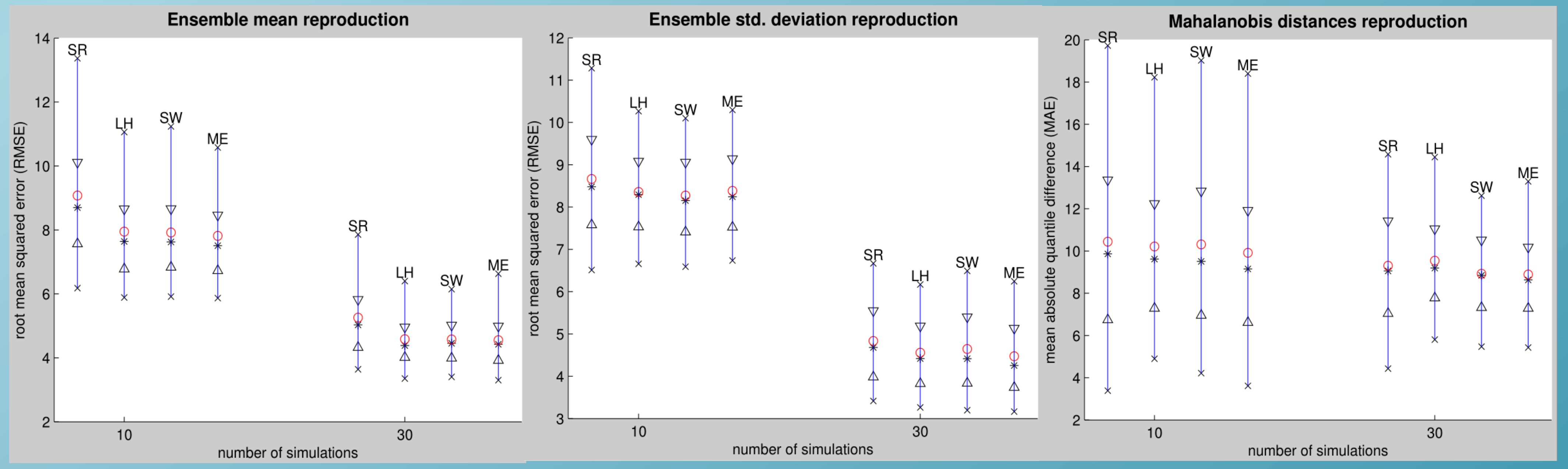

Figure 9. Reproduction of ensemble statistics i) mean, ii) std deviation and iii) mahalanobis distances, between concentration realizations and the ensemble average concentration field.

4. Conclusions and Discussion

The performance of the proposed ME sampling method was investigated in a hydrogeological context via a synthetic case study involving flow and transport in a heterogeneous porous medium, in comparison to stratified sampling methods LH and SL along with SR sampling. The statistics considered for hydraulic conductivity included the ensemble mean, standard deviation, as well as short-scale correlation and distribution of Mahalanobis distances from the ensemble mean. The reproduction of similar statistics for the ensemble concentration field resulting from solving a flow and transport boundary problem for each hydraulic conductivity realization, was also evaluated in the second part of the case study. For all statistics considered for both model inputs (Fig. 8) and outputs (Fig.9), ME sampling constitutes an equal if not more efficient simulation method than LH and SL sampling, as it can reproduce to a similar extent statistics of the reference conductivity and concentration fields, yet with smaller sampling variability than SR sampling. Concluding, the proposed ME sampling method offers a viable alternative to existing stratified sampling methods.

References

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

[2] 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**, 23–69.

[3] Switzer P. (2000): Multiple simulation of spatial fields. In: Heuvelink G, Lemmens M (eds) Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Coronet Books Inc., pp 629-635.

[4] Kyriakidis P. and Gaganis P. (2013): Efficient simulation of (log)normal random fields for hydrogeological applications. *Mathematical Geosciences*, **45**, 531–556.

[5] Hardin D. and Saff E. (2004): Discretizing manifolds via minimum energy points. Notices of the American Mathematical Society 51:1186,1194.

[6] Caers (2011): *Modeling Uncertainty in the Earth Sciences*, John Wiley & Sons.

[7] http://www.mathworks.com/matlabcentral/newsreader/view_thread/21747. Jason Bowman code distribution for Minimum Energy points on hyperspere.

GEOSTATENV