SAND2007-3658C

Integrating Sample Data with Ensemble Predictive Models for Efficient Contamination Event Characterization

6th Department of Homeland Security Conference on Chemical and Biological Technologies: Food, Restoration and Architectural Studies, June 5-8, 2007, Madison, WI

Sandia National Laboratories

Sean A. McKenna

Sandia National Laboratories, PO Box 5800 MS 0735, Albuquerque, NM, 87185-0735

samcken@sandia.gov, (505) 844-2450

Introduction

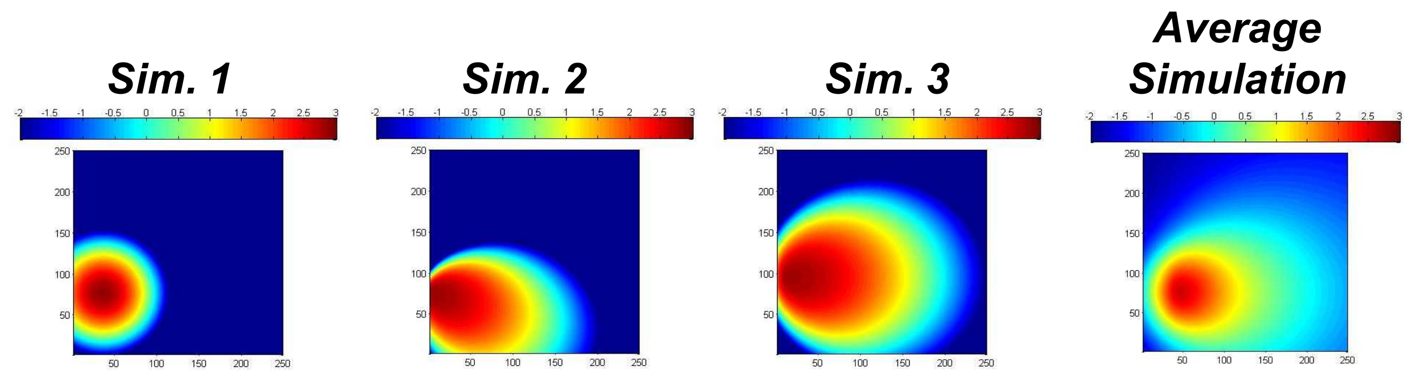
In the event of an aerosol contamination event, development of accurate and precise characterization of the magnitude and extent of the contamination is a key goal. This characterization provides multiple inputs to the response and recovery process including: source location identification, optimization of additional samples and prioritization of systems level recovery resources. This characterization is, ideally, the result of integrating as many different types of data sources as possible into a final, coherent characterization of the magnitude and extent of the contamination.

This study examines the ability of spatial statistical mapping tools to integrate relatively sparse sample data collected at point locations with an existing ensemble of numerical model predictions of the contamination event. The end result of this integration is definition of a non-parametric cdf of estimated concentration values at all locations. Monte Carlo sampling of the joint distribution of resultant cdfs provides expected case contamination, probability of exceeding prescribed concentration thresholds and posterior estimates of uncertainty. A demonstrative example using an ensemble of Gaussian plume simulations and limited point sampling is used to develop and evaluate an integration technique. Three different sets of results are compared: 1) Spatial estimation (Kriging) using only the limited sample data; 2) Estimates based only on a subset of the ensemble of model predictions that fit the sample data; and 3) The Colocated CoKriging (CoCoK) approach that integrates both sample data and the same subset of ensemble model results.

Example Problem

In this example problem, an aerosol release of a contaminant has occurred. An ensemble of 1000 Monte Carlo simulations of a numerical transport model exist and the ensemble of models adequately captures major uncertainties in the source location, amount of mass released, and meteorological conditions affecting transport and deposition. Each numerical model realization provides an estimate of the cumulative concentration over each 1x1km square within a 250x250km domain (62,400 cells). Here a simple Gaussian plume model (Diggle, 2003) is used to represent the aerosol transport model. Additionally, 35 samples of the concentration are located within this same domain.

Three example plume simulations and the average over all 1000 plumes are shown below

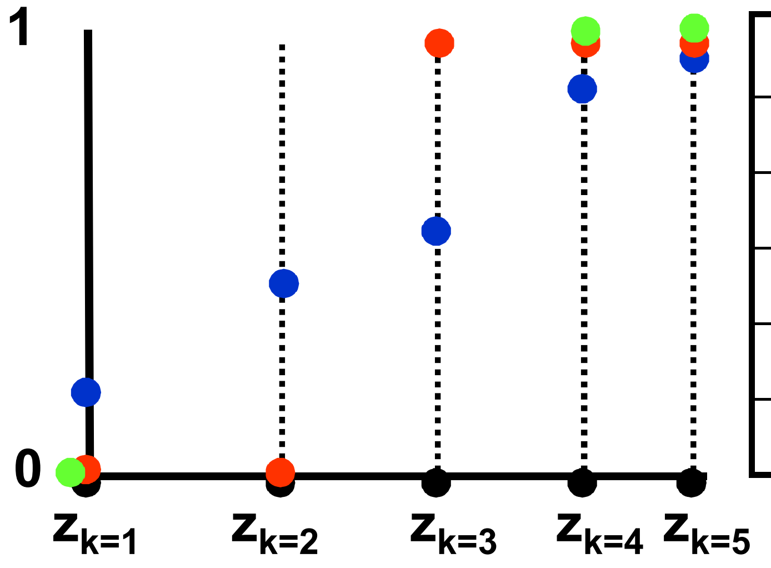


Color scale shows Log10 concentration values

Non-Parametric Indicator Coding

For every location, the cumulative distribution function (cdf) is estimated at a series of discrete concentration thresholds. Sample data are precise and transformed to 0 or 1 using the equation below (e.g., Sample 1). The proportion of the ensemble of simulations ≤ to each threshold provides continuous estimates of the cdf value at each threshold (e.g., Sample 2). Although not used here, data that only provide presence/absence information at discrete threshold values (e.g., Sample 3) can also be coded using this approach

$$i(\mathbf{x}_0; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{x}_0) \le z_k \\ 0 & \text{if } z(\mathbf{x}_0) > z_k \end{cases}$$



Threshold	Sample 1	Sample 2	Sample 3
$Z_{k=5}$	1.0	0.98	1.0
$Z_{k=4}$	1.0	0.91	1.0
$Z_{k=3}$	1.0	0.60	???
$Z_{k=2}$	0.0	0.35	???
$Z_{k=1}$	0.0	0.23	0.0

Data Integration with Colocated CoKriging

The kriging formulation provides an approach for unbiased, minimum variance linear estimation. Colocated Cokriging extends the kriging equations to incorporate a single covariate value at each estimation location, x_0. Here the cumulative probability at each threshold is estimated conditional to the n_s surrounding sample data and the n_m members of the simulation ensemble that are within some proximity measure of all sample data. The CoCoK estimate of the conditional probability is a weighted linear combination of the n_s indicators and the single covariate indicator value at the estimation location, $i^m(x_0)$, subject to the single unbiasedness constraint shown below

$$P(z(\mathbf{x_0}) \le z_k \mid s(1 \ldots n_s(\mathbf{x_0})), m(1 \ldots n_m) \mid S)) = \sum_{i=1}^{n_s(\mathbf{x})} \lambda_i^{\alpha} i^s(\mathbf{x}_i; z_k) + \lambda^{\beta}(\mathbf{x}_0; z_k) i^m(\mathbf{x}_0; z_k)$$

$$where: \sum_{i=1}^{n_s(\mathbf{x})} \lambda_i^{\alpha} + \lambda^{\beta}(\mathbf{x}_0) = 1.0$$

The weights, λ, are determined from solution of the kriging equations based on spatial covariances calculated from the sample data and correlation between the sample data and numerical simulations. If there are no numerical simulations, the λ_β weight is zero and the above equation resorts to the ordinary kriging system. The spatial covariances are derived from variograms calculated and modeled for each indicator threshold.

In the example problem, eight thresholds are defined in log10 concentration space as shown below. The first threshold, -2.0, corresponds to the detection limit of the sample data.

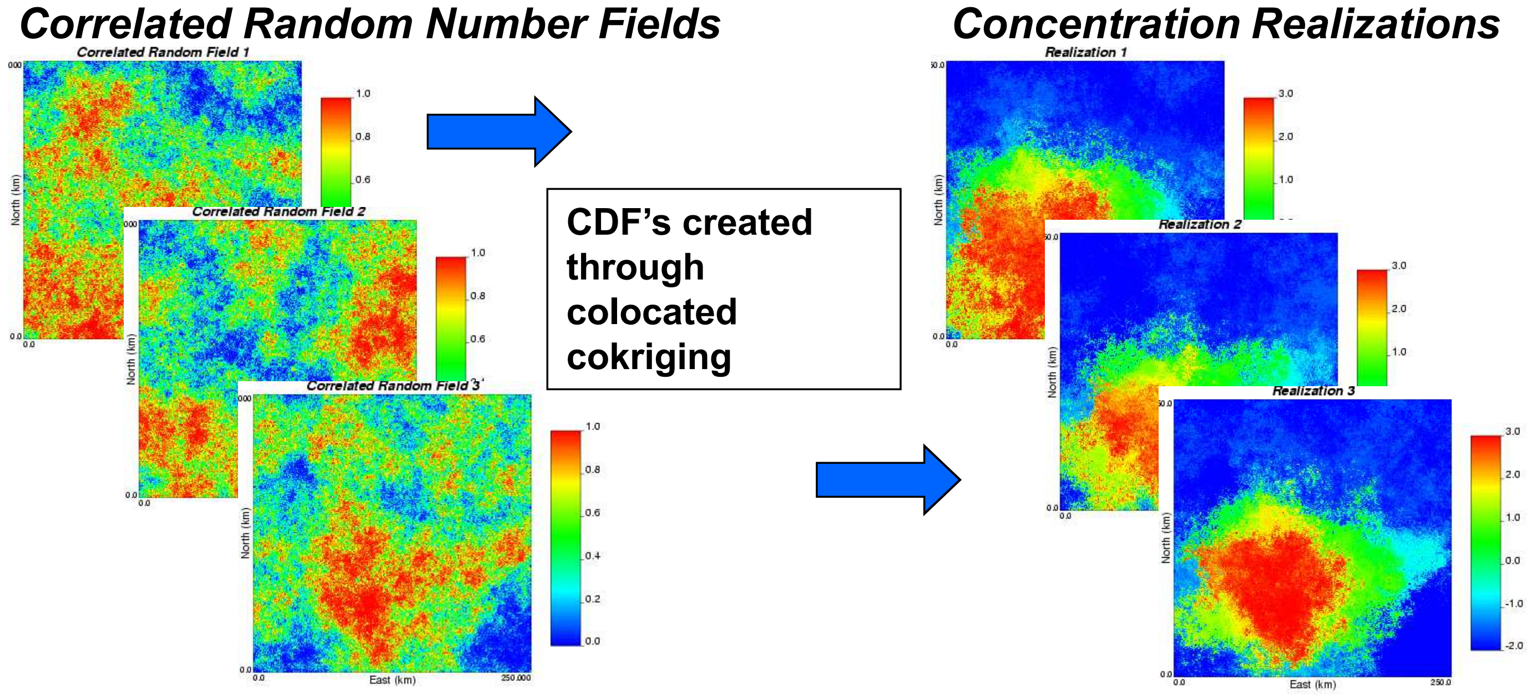
Thresholds: [-2.0, -1.85, -1.75, -1.55, -1.0, 0.5, 1.25, 2.15]

Sampling CDF Fields

Construction of non-parametric cdfs characterizes local uncertainty. Joint uncertainty is a function of multiple cfs, and the distributions are not independent :

$$P(z(u_j) \le z_k, j = 1, \ldots, J \mid (n_s, n_m)) \ne \prod_{j=1}^{J} F(u_j; z_c \mid (n_s, n_m))$$

cdf's are sampled simultaneously by applying a correlated field of random numbers. The value corresponding to each sampled point of the cdf is the concentration for that location. Multiple correlated random number fields produce multiple concentration realizations.

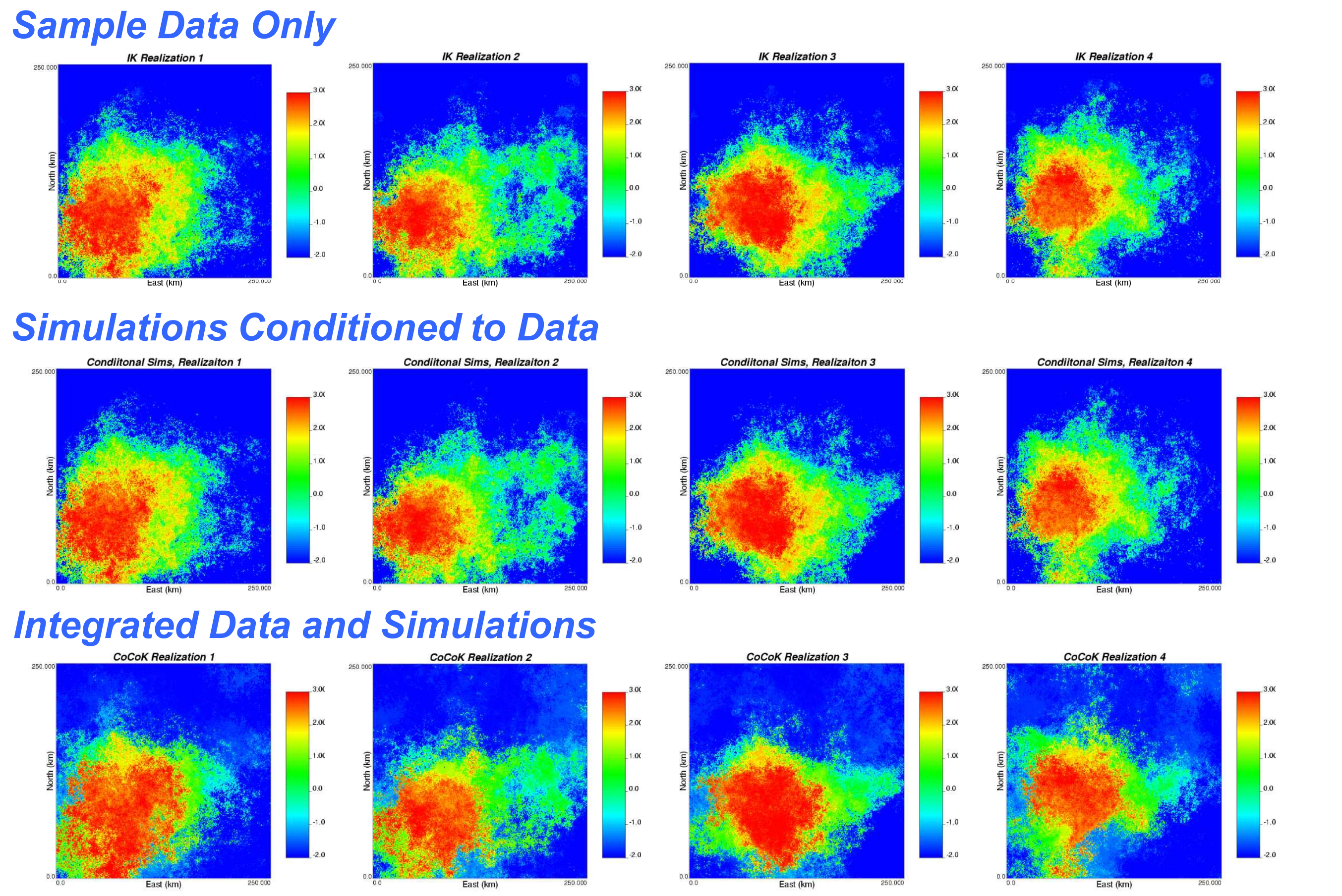

Correlated Random Number Fields

Concentration Realizations

CDF's created through colocated cokriging

Results: Example Simulations

Three sets of cdfs are created and then sampled 100 times: 1) Kriging of sample data; 2) 125 Gaussian plume simulations conditioned to the sample data; 3) Integrated Gaussian plume simulations and sample data

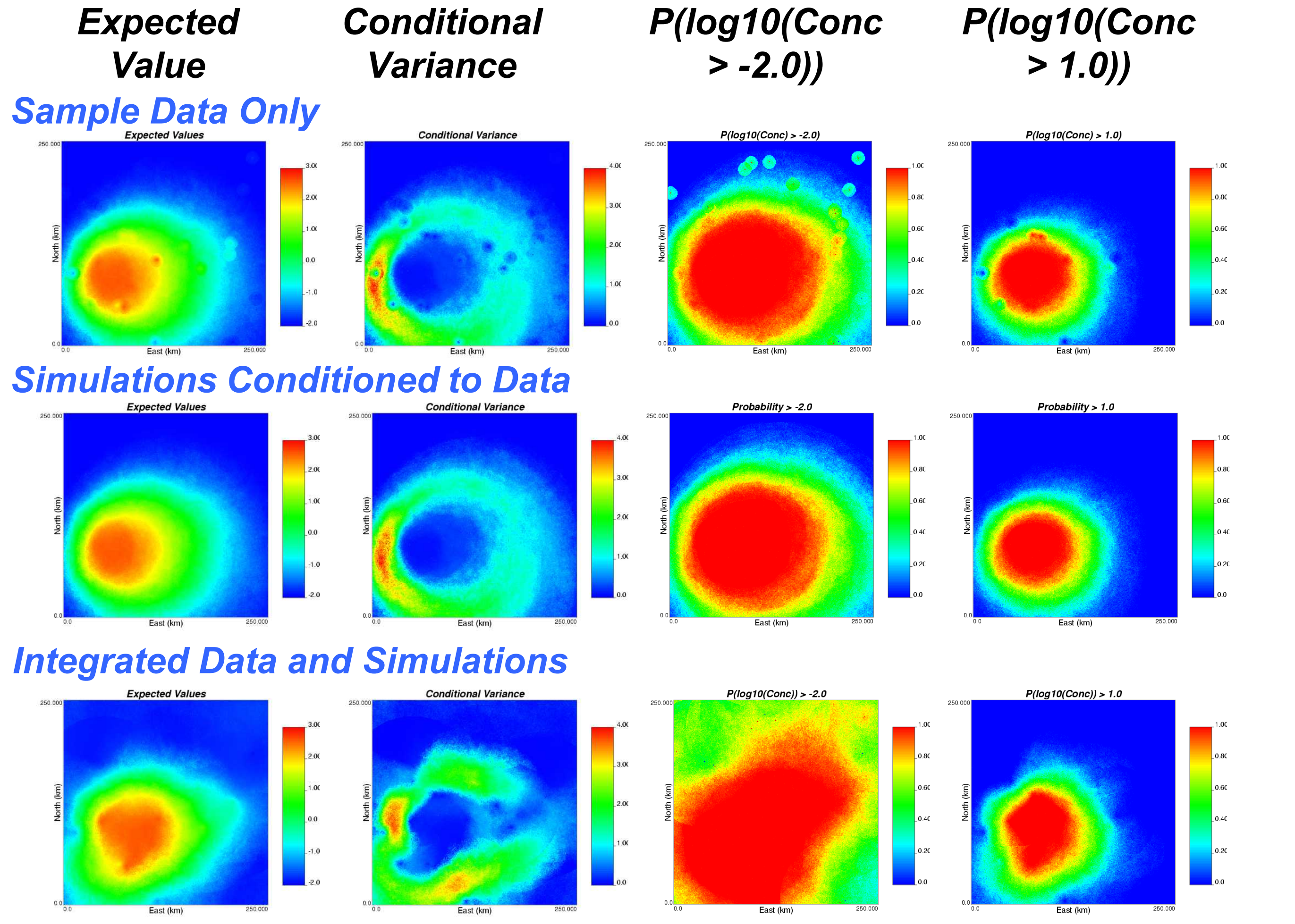

Sample Data Only

Simulations Conditioned to Data

Integrated Data and Simulations

Results: Uncertainty Mapping

From the 100 contaminant realizations, it is possible to map the average concentration field, the conditional variance of the concentration estimates (uncertainty), and the probability of exceeding different threshold concentration values.


Expected Value

Conditional Variance

P(log10(Conc > -2.0))

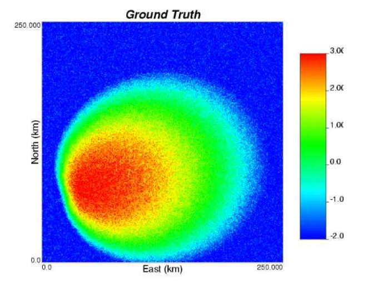
P(log10(Conc > 1.0))

Sample Data Only

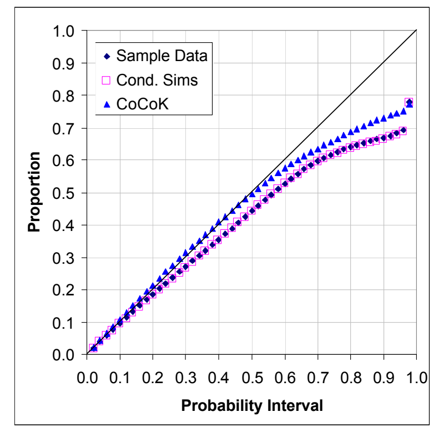
Simulations Conditioned to Data

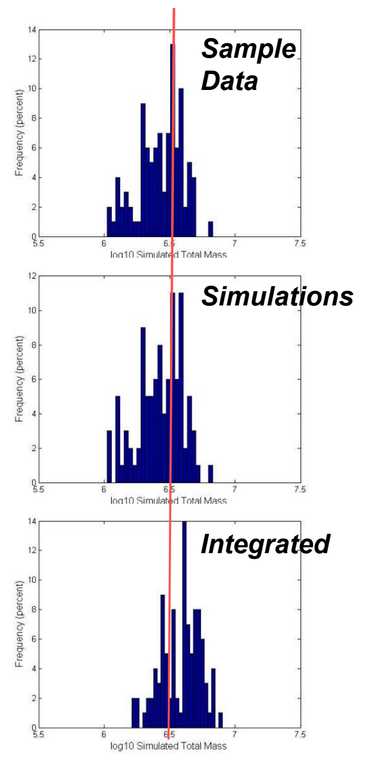
Integrated Data and Simulations

Results: Model Checking



The ground truth from which the samples were obtained is shown to the left and can be compared visually with the results shown above.





Probabilistic models are checked by determining the proportion of true values within each symmetric probability interval (image to right). Only those with proportion >= probability interval are accurate and this condition only occurs for the integrated sample/model results. None of the approaches adequately capture the widest probability intervals.

The total amount of mass released (log10(C)), is shown for each set of 100 realizations (histograms) and the ground truth value of 6.54 (red line) in the images above.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

http://www.sandia.gov/geostats