CMR/95-257

# DATA FUSION: A DECISION ANALYSIS TOOL THAT QUANTIFIES GEOLOGICAL AND PARAMETRIC UNCERTAINTY

David W. Porter, Ph.D
Thermo Electron Subsidiary
Coleman Research Corporation
Columbia, Maryland
(301) 621-8600

**ABSTRACT**

Engineering projects such as siting waste facilities and performing remediation are often driven by geological and hydrogeological uncertainties. Geological understanding and hydrogeological parameters such as hydraulic conductivity are needed to achieve reliable engineering design. Information form non-invasive and minimally invasive data sets offers potential for reduction in uncertainty, but a single data type does not usually meet all needs. Data Fusion uses Bayesian statistics to update prior knowledge with information from diverse data sets as the data is acquired. Prior knowledge takes the form of first principles models (e.g., groundwater flow) and spatial continuity models for heterogeneous properties. The variability of heterogeneous properties is modeled in a form motivated by statistical physics as a Markov random field. A computer reconstruction of targets of interest is produced within a quantified statistical uncertainty. The computed uncertainty provides a rational basis for identifying data gaps for assessing data worth to optimize data acquisition. Further, the computed uncertainty provides a way to determine the confidence of achieving adequate safety margins in engineering design. Beyond design, Data Fusion provides the basis for real time computer monitoring of remediation.

Working with the DOE Office of Technology (OTD), we have developed and patented a Data Fusion Workstation system that has been used on jobs at the Hanford, Savannah River, Pantex and Fernald DOE sites. Further, applications include an army depot at Letterkenney, PA and commercial industrial sites.

**INTRODUCTION**

Engineering projects involving hydrogeology are often driven by uncertainties. For activities such as environmental remediation or location of waste management facilities, cost effective solutions often depend on the confidence with which the hydrogeology is known. For example, to create a plume capture zone, answers are needed to questions about the number, depth, location and pumping schedules for purge wells. Computer simulation based on hydrogeological models provides answers. In theory, simulation can provide real-time monitoring of remediation so a plume can be "seen" as it is being cleaned up. But simulation is only as good as its geological and parametric inputs. The earth is very heterogeneous, and typical data sets are fragmented and disparate so there are substantial uncertainties.

Currently, environmental engineers do not have adequate tools to quantify uncertainty so they often rely solely on their judgement to build in sufficient safety margins. This tends to lead to overly conservative decisions that are often inordinately expensive. Data Fusion has value added as an engineering decision tool that quantifies uncertainty. In order to quantify uncertainty, fusion uses models in two different ways. First, models provide physical and statistical relationships between fragmented and disparate data sets, and fusion uses these relationships to extract geological and parametric information. Second, models are used in computer simulation of remediation (e.g., to track plume movement) where the simulation is based on geological and parametric inputs. Data Fusion and modeling will become even more important as new technology provides additional data sources to describe the sites subsurface materials and the pollutants that may be passing through them. This has already happened in numerical weather prediction and physical oceanography.

With reduced cleanup budgets and ever increasing movement towards more comprehensive risk assessment with predictive models having quantified uncertainties, Data Fusion and modeling are technologies for the times. Using Data Fusion and

## DISCLAIMER

## DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

modeling, decision makers have a quantitative basis for action so the following benefits can be realized:

- Enables remedial simulation to optimize cleanup. Remediation solutions can be exercised in the computer to match quantified safety margins.
- Enables real-time monitoring during remediation. Contaminant plumes can be continuously monitored while they are being cleaned up.
- Provides quantitative basis for cost reduction/avoidance.
- Establishes data worth, before expending funds for field data acquisition, to determine if reduction in uncertainty pays for the cost of acquisition.
- Derives the most out of existing data sets to avoid cost of unnecessary acquisition.

Data Fusion and modeling have a solid foundation in the hydrogeological community. Freeze et.al. published a framework for hydrogeological decision analysis in references 1 to 4. A pragmatic engineering approach to decision making is described that balances benefits, costs, and uncertainties. We have adopted the decision analysis viewpoint and approach in our Data Fusion as shown in Figure 1.1. Engineers face uncertainty in parameters (such as hydraulic conductivity) and in the geometry of the problem through the geology. Data Fusion quantifies geological and parametric uncertainty. As shown in Figure 1.1, hydrogeological simulation is performed (e.g., to see plume movement) using geological and parametric inputs. Fusion propagates geological and parametric uncertainties through the simulation so the confidence in plume movement is quantified. Engineering reliability uncertainties in the engineered components of remediation also enter into decisions, but the hydrogeological uncertainties usually dominate.

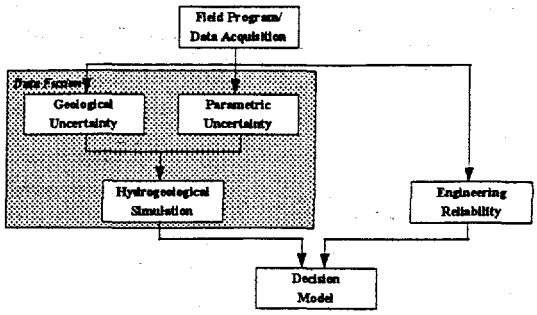


**Figure 1.1 Data Fusion Role in Decision Analysis**

## 2. THEORY

A Data Fusion perspective is presented, beginning with the hydrogeological foundation. Data assimilation is described as a starting point for fusion. Then the following fusion methods are described: Markov Random Field (MRF) model, Square Root Information Smoother (SRIS), and Generalized Expectation Maximization (GEM) method.

The methodology of references 1 to 4 views hydrogeology as a predictive science that must incorporate the fundamental heterogeneity of the subsurface. Consequently, hydraulic conductivity is treated as an autocorrelated spatial stochastic process so the variability and spatial continuity of the conductivity is modeled.

Bayesian estimation is performed, but Freeze et al. point out that a limitation of the Bayesian approach is the application of inverse modeling. Data assimilation methods incorporate inverse modeling in a fundamental way, but they are too numerically demanding for practical applications (see Ref. 5). Consequently, it is not practical to combine all the important data sets to reduce and quantify the uncertainty in the subsurface. Data Fusion resolves this limitation by building on data assimilation to produce a full inverse modeling approach that is numerically practical.

Data assimilation methods are well established in numerical weather prediction, have moved into physical oceanography and are being established in hydrogeology (see Refs. 5 to 8). The methods take many forms from adjoint to variational to Kalman filtering.

Our Data Fusion approach provides Bayesian inverse modeling as shown in Figure 2.1. It begins with prior knowledge about the state variables to be

estimated in the form of first principles models, spatial continuity models in the form of spatial stochastic processes, and uncertain initial conditions. Fusion performs Bayesian updates using measured data and data models as the data is acquired. Posterior state knowledge is produced in the form of state estimates with quantified uncertainties. Residual model fit errors are used as diagnostics to detect discontinuities, perform data validation, and to tune prior statistics such as spatial correlation distances and standard deviations.

Our methods are mathematically equivalent to the Kalman filter. However, we represent spatial stochastic processes using a Markov Random Field (MRF) borrowed from statistical physics (see Ref. 9). By generalizing the Square Root Information Smoother (SRIS) presented by Bierman (Ref. 10), the MRF is incorporated to produce a numerically practical solution. A complete theoretical development of the Data Fusion theory with a description of the Data Fusion System (DFS) software architecture is found in Reference 11.

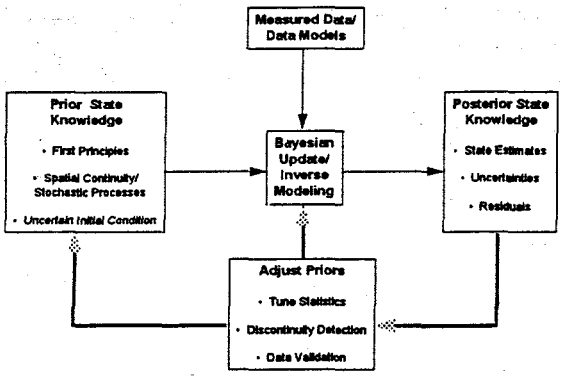


**Figure 2.1 Data Fusion Modeling Uses Bayesian Statistics**

**Data Assimilation**

McLaughlin points out in Reference 5 that many of the data assimilation techniques can be viewed as special cases of the Kalman filter. The Kalman filter provides Bayesian inverse modeling viewing heterogeneities as spatial stochastic processes as needed by the hydrogeological decision analysis methodology. The Kalman filter also includes a model error noise to account for approximation errors in the first principles models used for inverse modeling. Consequently, the Kalman filter produces a solution that honors the data within measurement error noise, the first principles models within model error noise, and spatial continuity within the spatial autocorrelation.

The inclusion of model noise is important to achieve predictive modeling with quantified uncertainty. Applications in a variety of communities have shown the necessity of recognizing that models have noise just like data. For example, satellite orbit determination filters required ad hoc fix up procedures before model noise was used to account for solar wind and gravity disturbances (see Reference 10). Sensor mixing for inertial navigation uses model noise for inertial instrument drift and environmental disturbances like gravity errors and, in marine applications, ocean current disturbances (see Reference 10).

The inverse problem can be formulated in a geostatistical or indirect iterative manner as shown in Reference 8. Once statistical correlations are established between measurements and variables to be estimated, the geostatistical approach computes the best linear estimate in one step. The indirect iterative approach iteratively minimizes a least squares penalty function that penalizes data/model mismatches, excessive model error and, excessive variability in heterogeneous variables. Reference 8 shows that the geostatistical approach is mathematically equivalent to one step of the iterative approach. By iterating to convergence, the iterative approach can provide better estimates and better quantification of uncertainties because it is a nonlinear estimator not subject to the linear restriction of the geostatistical approach.

The difficulty with current data assimilation methods is that they are too numerically demanding for practical application. We have formulated the Data Fusion to be numerically practical and to retain the desirable features of the data assimilation methods. In fact, Data Fusion can be viewed as a data assimilation method that is numerically practical in today's UNIX workstations or Pentium class PCs.

**Data Fusion**

Our methods are mathematically equivalent to the Kalman filter, using the Square Root Information Smoother (SRIS) to produce a numerically practical solution. Model noise is incorporated in order to provide a complete predictive modeling capability. Data Fusion is formulated as an indirect iterative method to achieve the best state estimates and quantification of uncertainty.

The key to achieving a numerically practical approach is to return to the basics of spatial stochastic processes. The Kalman filter has difficulty

with spatial processing because it uses a causality property to break large processing problems down into a sequence of smaller problems. Causality means that there is a past causing a future. Causality is a powerful property for processes that evolve in time where there is a past and a future. But causality does not work in space.

The Kalman filter was formulated as a generalization of the Wiener filter to incorporate physical models and to be in a form more suitable for computer implementation. But the Wiener filter does not require causality so causality is not an inherent restriction. Wiener's original work and the field of statistical physics are closely tied together. It is through statistical physics that we find the replacement for causality in the concept of a Markov Random Field (MRF) to make computations practical.

**Markov Random Field (MRF)**

The connection between Bayesian estimation and statistical physics MRF ideas was made in the computer vision community in References 9, 12 and 13. The MRF provides a way to model large scale statistical structure using only local computations. This means that large unmanageable spatial processing problems can be broken down into smaller local problems that are practical to compute.

MRF models are used in statistical physics for such problems as chemical annealing to determine lowest energy states. This has a direct analogy to Bayesian inverse modeling in determining the minimum value of a least squares penalty function for the indirect iterative method. In computer vision, MRF methods are used as the basis for Data Fusion solutions for stereo vision, shape determination from shading data, tracking object motion, and tomographic image reconstruction. However, computer vision uses the technique of stochastic relaxation to do the actual computing, but we use the SRIS technique.

**Square Root Information Smoother (SRIS)**

We start with the representation of an MRF as a spatial autoregression (see Reference 14). The autoregression has a local computational form that expresses the MRF as an interpolation of nearby values plus an interpolation error that is uncorrelated over space.

The autoregression puts the MRF in a form compatible with the data equation idea used by Bierman for the SRIS (in Reference 10). Bierman used the data equation to express prior knowledge on first principles models and statistical correlations as pseudo-data as if it were just more data for doing Bayesian estimation. Since the pseudo-data and data models are all local in space, the processing can be broken down sequentially in space so it becomes practical.

The SRIS has many possible forms depending on the specific details of how data and pseudo-data are represented. In fact, the SRIS is actually a family of algorithms that can be designed according to a set of information principles (see Reference 15). The nonlinear iteration to produce the indirect iterative solution is provided by the Trust Region method for numerical optimization (see Reference 16) with the option of a Gauss-Newton step halving method.

**Generalized Expectation Maximization (GEM) Method**

The GEM method provides a practical means to tune statistical parameters in order to adjust prior knowledge based on the data themselves as shown in Figure 2.1. GEM is mathematically equivalent to the likelihood approach for estimating statistical parameters described in Reference 8. The approach used to calibrate system noise for groundwater simulations by Van Geer, et. al. in Reference 17 is similar in concept to GEM. But GEM is much more flexible and provides statistically optimal maximum likelihood estimates for statistical parameters such as spatial correlation distance and standard deviation.

GEM views the measured data as incomplete data for the purpose of estimating the statistical parameters (see Reference 18). A complete data set is specified from which the parameters could easily have been estimated if the complete data had been measured. Then a sufficient statistic of the complete data is estimated in an expectation step based on starting values for the parameters being tuned. The parameters are updated based on the sufficient statistic in a likelihood maximization step and the process is iterated. GEM has desirable statistical convergence properties, has been used in a host of agricultural, economic and scientific applications (see Reference 18), and used for military, scientific and image processing applications (See Reference 19).

## HANFORD APPLICATION

Data Fusion modeling was applied at the Hanford 200 West Area. The objective was to map a thin caliche layer 30 to 40 meters below the surface. Scattered well data with caliche picks was provided by the Hanford site. In order to obtain a more

detailed characterization, a geophysical survey was conducted to obtain densely spaced non-intrusive data. The survey design consisted of four seismic reflection and refraction lines which are shown in Figure 3.1. Figure 3.1 shows the Data Fusion results using the seismic measurements and well picks in the Hanford 200 West area. The top of the caliche layer, a possible barrier to contaminants, is shown at the bottom of the chair cut. Top of caliche picks were obtained from 25 wells. The individual interpretations for seismic reflection data, without checkshot information, provide almost no information on the caliche. Using reflection data alone, inadequate seismic velocity information was available to interpret travel times as depths to the caliche. The caliche did not support a refraction so no depth information was available from the refraction measurements. Data Fusion by jointly processing all of the data uses seismic velocity information from the refraction measurements to interpret the reflection travel times as depths. This demonstrates the value added by the ability of Data Fusion to produce a complete picture from multiple sensors that individually have only partial visibility of a subsurface feature.

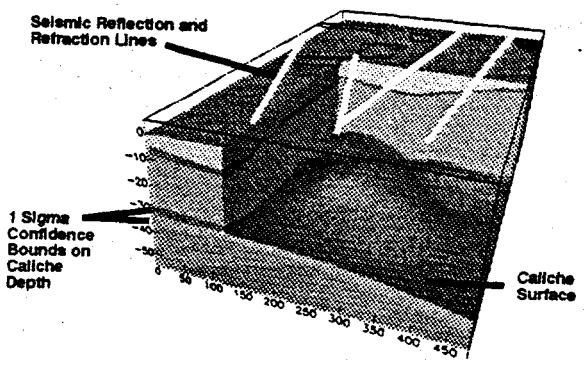


**Figure 3.1 Hanford 200 West Site Caliche Delineation**

## 4. PANTEX APPLICATION

A groundwater model was calibrated using Data Fusion modeling for a perched aquifer under Zone 12 at Pantex using hydraulic head data, slug tests, and recharge information. A steady-state finite-difference model was used that is similar to the USGS MODFLOW model and is called MODLIKE. A good model fit was achieved with an RMS head residual of only .4 feet. Hydraulic conductivity heterogeneity was estimated in order to provide flow pathlines shown in figure 4.1 with confidence within the region of data coverage.

As the conceptual model was modified to improve results, fusion converged rapidly for each conceptual model. Generally it took 8 or 9 numerical iterations at approximately 15 minutes of Indigo II time. This means that fusion modeling is fast enough to be used for real time field model updating. This provides a capability that has never before been possible with conventional manual model calibration.

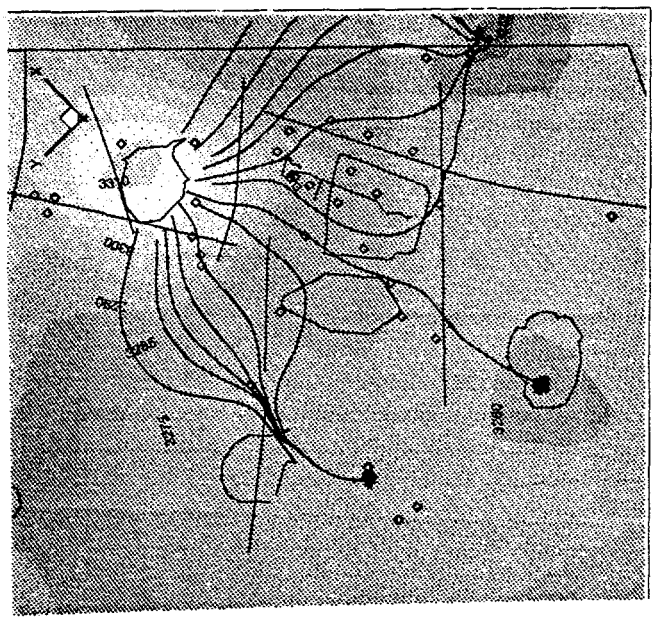

**Figure 4.1 Head Contours (ft) And Pathlines**

## REFERENCES


1. Freeze, R.A., Massman J., Smith, L., Sperling, T. and James, B., "Hydrogeological Decision Analysis: 1. A Framework, " Ground Water, Sep-Oct. 1990.
2. Massman, J., Freeze, R.A. J., Smith, L., Sperling, T. and James, B., "Hydrogeological Decision Analysis: 2. Applications to Ground-Water Contamination, "Ground Water, July-Aug. 1991.
3. Sperling, T., Freeze, R.A. Massman, J., Smith, L. and James, B., "hydrogeological Decision Analysis: 3. Application to Design of a Ground-Water Control Systems at an Open Pit Mine," Ground Water, May-June. 1992.
4. Freeze, R.A., Massman, J., Smith, L., Sperling, T. and James, B., " Hydrogeological Decision Analysis: 4. The concept of Data Worth and its Use in the development of Site Investigation Strategies," Ground Water, Jul-Aug. 1992.
5. McLaughlin, D., " Recent Developments in Hydrologic Data Assimilation," Prepared for the U.S. Report to the IUGG, Jun. 1994.

6. Ghil, m., 1989, "Meteorological Data Assimilation for Oceanographers, Part 1, Description and Theoretical Framework, "Dynamics of Atmospheres and Oceans, 13, pp. 171-218.
7. Coutier, P., Derber, J., Errico, R., Louis, J.F. and Vukicevic, T., 1993, "Important Literature on the use of Adjoint, Variational Methods and the Kalman Filter in Meteorology, " Tellus, 45A, pp. 342-357.
8. Carrera, J., and Glorioso, L., 1991, "On Geostatistical Formulations of the Ground Water Flow Inverse Problem," Advances in Water Resources, 14, No. 5, pp 273-283.
9. Chellappa, R. and Jain, A., Markov Random Fields: Theory and Application, Boston:Academic Press, 1991.
10. Bierman, G. J., "Factorization Methods for Discrete Sequential Estimation," New York: Academic Press. 1977.
11. "Geophysical Data Fusion for Subsurface Imaging," Morgantown Energy Technology Center, Contract # DE-AC21-92MC29106, Coleman Research Corp. Final Report, June 1995.
12. Geman, S., and D. Geman, "Stochastic Relaxation, Gibbs Distribution and the Bayesain Restoration of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, pp 721-741, 1984.
13. Clark, J.J. and A. L. Yuille, Data Fusion for Sensory Information Processing Systems, Kluwer-Academic Publishers, Boston 1990.
14. Whittle, P., "On Stationary Processes in the Plane," Biometrika, Vol. 41, Dec. 1954.
15. Porter, D.W., "Quantitive Data Fusion: A Distributed/Parallel Approach to Surveillance, Tracking, and Navagation using Information Filtering," Fifth Joint Data Fusion Symposium, Johns Hopkins University/Applied Physics Laboratory, October 1991.
16. Vandergraft, J.S., "Efficient Optimization Methods for Maximum Likelihood Parameter Estimation," Proceedings of 24th Conference on Decision and Control, Ft. Lauderdale, FL December 1985.
17. VanGeer, F.C., C.B.M. Te Stoet, and Z. Yangxiao, "Using Kalman Filtering to Improve and Quantify the Uncertainty of Numerical Ground Water Simulations, 1. The role of System Noise and its Calibration," Water Resource Res., 27(8), pgs. 1987-1994, 1991.
18. Dempster, A.P., N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm, " J. Royal Statistical Society, Ser. B.39, 1977.
19. Levy L.J., and D.W. Porter, "Large-Scale System Performance Prediction with Confidence from Limited Field Testing Using Parameter Identification," the John Hopkins APL Technical Digest, Vol. 13, No. 2, 1992.