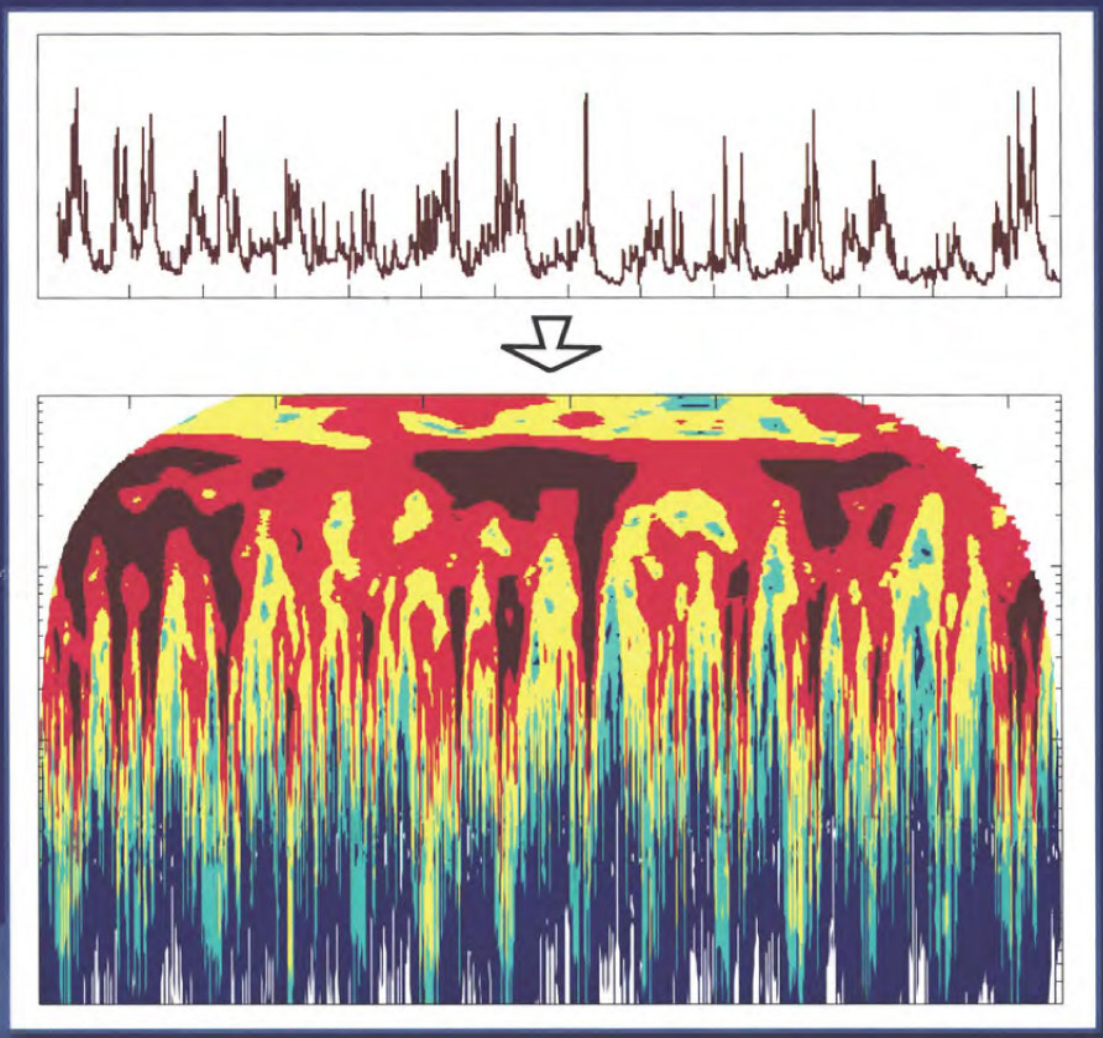
Subsurface Hydrology

Data Integration for Properties and Processes



David W. Hyndman, Frederick D. Day-Lewis,
and Kamini Singha
Editors

AGU

Geophysical Monograph Series

Including
IUGG Volumes
Maurice Ewing Volumes
Mineral Physics Volumes

Geophysical Monograph Series

134 **The North Atlantic Oscillation: Climatic Significance and Environmental Impact** *James W. Hurrell, Yochanan Kushnir, Geir Ottersen, and Martin Visbeck (Eds.)*

135 **Prediction in Geomorphology** *Peter R. Wilcock and Richard M. Iverson (Eds.)*

136 **The Central Atlantic Magmatic Province: Insights from Fragments of Pangea** *W. Hames, J. G. McHone, P. Renne, and C. Ruppel (Eds.)*

137 **Earth's Climate and Orbital Eccentricity: The Marine Isotope Stage 11 Question** *André W. Droxler, Richard Z. Poore, and Lloyd H. Burckle (Eds.)*

138 **Inside the Subduction Factory** *John Eiler (Ed.)*

139 **Volcanism and the Earth's Atmosphere** *Alan Robock and Clive Oppenheimer (Eds.)*

140 **Explosive Subaqueous Volcanism** *James D. L. White, John L. Smellie, and David A. Clague (Eds.)*

141 **Solar Variability and Its Effects on Climate** *Judit M. Pap and Peter Fox (Eds.)*

142 **Disturbances in Geospace: The Storm-Substorm Relationship** *A. Surjalal Sharma, Yohsuke Kamide, and Gurbax S. Lakhima (Eds.)*

143 **Mt. Etna: Volcano Laboratory** *Alessandro Bonaccorso, Sonia Calvari, Mauro Coltelli, Ciro Del Negro, and Susanna Falsaperla (Eds.)*

144 **The Subseafloor Biosphere at Mid-Ocean Ridges** *William S. D. Wilcock, Edward F. DeLong, Deborah S. Kelley, John A. Baross, and S. Craig Cary (Eds.)*

145 **Timescales of the Paleomagnetic Field** *James E. T. Channell, Dennis V. Kent, William Lowrie, and Joseph G. Meert (Eds.)*

146 **The Extreme Proterozoic: Geology, Geochemistry, and Climate** *Gregory S. Jenkins, Mark A. S. McMenamin, Christopher P. McKay, and Linda Sohl (Eds.)*

147 **Earth's Climate: The Ocean–Atmosphere Interaction** *Chunzai Wang, Shang-Ping Xie, and James A. Carton (Eds.)*

148 **Mid-Ocean Ridges: Hydrothermal Interactions Between the Lithosphere and Oceans** *Christopher R. German, Jian Lin, and Lindsay M. Parson (Eds.)*

149 **Continent-Ocean Interactions Within East Asian Marginal Seas** *Peter Clift, Wolfgang Kuhnt, Pinxian Wang, and Dennis Hayes (Eds.)*

150 **The State of the Planet: Frontiers and Challenges in-Geophysics** *Robert Stephen John Sparks and Christopher John Hawkesworth (Eds.)*

151 **The Cenozoic Southern Ocean: Tectonics, Sedimentation, and Climate Change Between Australia and Antarctica** *Neville Exon, James P. Kennett, and Mitchell Malone (Eds.)*

152 **Sea Salt Aerosol Production: Mechanisms, Methods, Measurements, and Models** *Ernie R. Lewis and Stephen E. Schwartz*

153 **Ecosystems and Land Use Change** *Ruth S. DeFries, Gregory P. Anser, and Richard A. Houghton (Eds.)*

154 **The Rocky Mountain Region—An Evolving Lithosphere: Tectonics, Geochemistry, and Geophysics** *Karl E. Karlstrom and G. Randy Keller (Eds.)*

155 **The Inner Magnetosphere: Physics and Modeling** *Tuija I. Pulkkinen, Nikolai A. Tsyganenko, and Reiner H. W. Friedel (Eds.)*

156 **Particle Acceleration in Astrophysical Plasmas: Geospace and Beyond** *Dennis Gallagher, James Horwitz, Joseph Perez, Robert Preece, and John Quenby (Eds.)*

157 **Seismic Earth: Array Analysis of Broadband Seismograms** *Alan Levander and Guust Nolet (Eds.)*

158 **The Nordic Seas: An Integrated Perspective** *Helge Drange, Trond Dokken, Tore Furevik, Rüdiger Gerdes, and Wolfgang Berger (Eds.)*

159 **Inner Magnetosphere Interactions: New Perspectives From Imaging** *James Burch, Michael Schulz, and Harlan Spence (Eds.)*

160 **Earth's Deep Mantle: Structure, Composition, and Evolution** *Robert D. van der Hilst, Jay D. Bass, Jan Matas, and Jeannot Trampert (Eds.)*

161 **Circulation in the Gulf of Mexico: Observations and Models** *Wilton Sturges and Alexis Lugo-Fernandez (Eds.)*

162 **Dynamics of Fluids and Transport Through Fractured Rock** *Boris Faybishenko, Paul A. Witherspoon, and John Gale (Eds.)*

163 **Remote Sensing of Northern Hydrology: Measuring Environmental Change** *Claude R. Duguay and Alain Pietroniro (Eds.)*

164 **Archean Geodynamics and Environments** *Keith Benn, Jean-Claude Mareschal, and Kent C. Condie (Eds.)*

165 **Solar Eruptions and Energetic Particles** *Natchimuthukonar Gopalswamy, Richard Mewaldt, and Jarmo Torsti (Eds.)*

166 **Back-Arc Spreading Systems: Geological, Biological, Chemical, and Physical Interactions** *David M. Christie, Charles Fisher, Sang-Mook Lee, and Sharon Givens (Eds.)*

167 **Recurrent Magnetic Storms: Corotating Solar Wind Streams** *Bruce Tsurutani, Robert McPherron, Walter Gonzalez, Gang Lu, José H. A. Sobral, and Natchimuthukonar Gopalswamy (Eds.)*

168 **Earth's Deep Water Cycle** *Steven D. Jacobsen and Suzan van der Lee (Eds.)*

169 **Magnetic ULF Waves: Synthesis and New Directions** *Kazue Takahashi, Peter J. Chi, Richard E. Denton, and Robert L. Lysak (Eds.)*

170 **Earthquakes: Radiated Energy and the Physics of Faulting** *Rachel Abercrombie, Art McGarr, Hiroo Kanamori, and Giulio Di Toro (Eds.)*

Geophysical Monograph 171

Subsurface Hydrology: Data Integration for Properties and Processes

David W. Hyndman
Frederick D. Day-Lewis
Kamini Singha
Editors

American Geophysical Union
Washington, DC

Library of Congress Cataloging-in-Publication Data

Subsurface hydrology : data integration for properties and processes / David W. Hyndman, Frederick D. Day-Lewis, Kamini Singha, editors.
p. cm. -- (Geophysical monograph ; 171)
ISBN 978-0-87590-437-5
1. Groundwater flow--Mathematical models. I. Hyndman, David W. II. Day-Lewis, Frederick D. III. Singha, Kamini. IV. American Geophysical Union.
GB1197.7.S84 2007
551.49--dc22
2007017693

ISBN 978-0-87590-437-5

ISSN 0065-8448

Front cover image: Spectral analysis of the stream discharge hydrograph (top) for the Muskegon River in central-northern Michigan, USA, reveals a rich time-varying power spectrum (bottom). Direct comparison of the discharge power spectrum to that of precipitation or water table fluctuations can provide significant insight into watershed processes. *Courtesy of David W. Hyndman*.

Printed in the United States of America.

CONTENTS

PREFACE

Groundwater is the principal source of drinking water for over 1.5 billion people. With increasing demands for potable water, continued threats to water quality, and growing concerns about climate change, the processes controlling groundwater availability are of paramount concern. There are also considerable concerns about the sustainability of groundwater supplies, given that much of the water withdrawn from aquifers today was recharged thousands of years ago. Data about hydrologic properties controlling flow and transport are needed to predict and simulate water-resources management practices, aquifer remediation, well-head protection, ecosystem management, and geologic isolation of radioactive waste. As the study of fundamental processes moves forward, we find that the physical processes of flow are complex at all scales, and furthermore are coupled with chemical and biological processes. In the 21st century, hydrologic scientists increasingly find themselves considering a diverse range of processes, data types, and analytical tools to help unravel processes controlling subsurface dynamics.

Quantifying the nature of hydrogeologic processes such as fluid flow, contaminant transport, or groundwater-surface-water interactions is difficult due to poor spatial sampling, heterogeneity at multiple scales, and time-varying properties. This book provides a series of examples where multiple data types have been integrated to better understand subsurface hydrology. We hope it serves to stimulate discussion and research on ways to improve our understanding on hydrologic processes, which are increasingly relevant as societal needs for clean water become more pressing. We thank the authors and reviewers of the chapters contained within this monograph and Allan Graubard, our AGU acquisitions editor.

David W. Hyndman
Frederick D. Day-Lewis
Kamini Singha
Editors

Subsurface Hydrology: Data Integration for Properties and Processes
Geophysical Monograph Series 171

10.1029/171GM01

INTRODUCTION

Kamini Singha

Department of Geosciences, The Pennsylvania State University, University Park, Pennsylvania, USA

David W. Hyndman

Department of Geological Sciences, Michigan State University, Michigan, USA

Frederick D. Day-Lewis

U.S. Geological Survey, Office of Ground Water, Bureau of Geophysics, Storrs, Connecticut, USA

Understanding the processes that control water movement in the subsurface has been recognized as a "grand challenge" in environmental science [*National Research Council*, 2001b]. Research into methods to estimate hydrologic parameters that control water movement extends at least back to *Theis* [1935], who worked simultaneously on methods to predict (forward model) aquifer response to pumping, and also to estimate (using an inverse model) the controlling hydrologic parameters—transmissivity and storativity. Seventy years after Theis' pioneering work, hydrologists continue to use pumping tests and slug tests to characterize heterogeneous aquifers. Despite advances in modeling tools and inverse methods, aquifer characterization remains an extremely difficult problem due to spatial heterogeneity, temporal variability, and coupling between chemical, physical, and biological processes.

The concept of data integration (also called data fusion or data assimilation) involves merging multiple data types to develop more reliable predictive models, and to answer basic and applied science questions. In many applications, combinations of complementary data types has been shown to yield more information than analysis of more abundant data of a single type [*National Research Council*, 2000; *National Research Council*, 2001a]. Ideally, this would involve a seamless connection of field data across broad ranges of data types, temporal scales, and spatial scales from pores to watersheds and beyond. In practice, hydrologic measurements tend to be either sparse, local, and representative of only small volumes of the subsurface, or integrated over large volumes making it difficult to characterize heterogeneous hydrologic parameters. As a result, there remains a need for cost-effective data sources, and novel approaches to integrate multiple data types that consider coupled processes across multiple scales. Data integration is thus critical to improve our understanding of complex, multi-scale hydrologic processes, which often have feedbacks with other physical, chemical, and biological processes at multiple scales.

Reliable predictions of future system behavior depend on our ability to develop models that accurately represent field conditions based on collected data, while simulating key processes with a sparse set of parameters. With limited data, the problem of model identification is generally poorly constrained; as additional data types are considered, however, the intersection between viable sets of models becomes smaller (Figure 1) and estimates of parameters and rates of processes in the field improve. Recognition of this synergy is evidenced by the increasing number of integrated analyses of multiple data types, and a growing realization that simultaneous consideration of multiple data types, provides improved ways to characterize and monitor subsurface hydrologic properties and processes [e.g., *Hubbard and Hornberger*, 2006].

There are a wide range of data types that can be used to improve our understanding of hydrologic processes, ranging from direct estimates of hydrologic parameters (e.g., perme-

Subsurface Hydrology: Data Integration for Properties and Processes
Geophysical Monograph Series 171

10.1029/171GM02

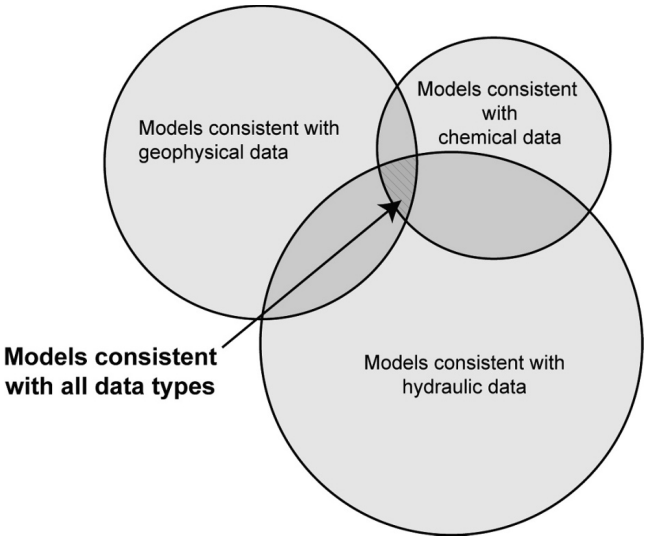


Figure 1. Sets of models can explain different data types. The intersection identifies the "best" model or models that represent the system across the integrated range of available data types.

ameter measurements on cores or flowmeter measurements of hydraulic conductivity) to indirect information from geologic maps, geophysical tomography, or quantities related to parameters of interest through physical models such as heat or solute transport. Table 1 provides a list of representative references where the listed data type is used to estimate parameters in subsurface models. This list is by no means exhaustive, but indicates the diversity of information sources used in hydrology. While data integration is increasingly implemented in hydrologic studies, it is also an active area of research due to the complexities of scale and measurement support volume, data weighting, model parameterization, realistic representation of geology in numerical models, and implementation of coupled-process numerical models.

This volume provides a broad sampling of papers that represent the current state of the science of data integration for subsurface hydrology. The premise underlying the collected work in this volume is that simultaneous consideration of multiple data types allows for an improved understanding of subsurface hydrology. The monograph is divided into four sections: (1) approaches to quantitative data integration; (2) data integration for characterization of hydrologic properties; (3) data integration for understanding hydrologic processes; and (4) meta analysis.

The first section includes papers on approaches to hydrologic data integration, which range from qualitative interpretation of multiple data types to rigorous non-linear inversion of coupled-process numerical models. In the last few decades, non-linear regression models that estimate subsurface properties based on groundwater data [e.g., *Neuman and Yakowitz*, 1979; *Gorelick*, 1990; *Gailey et al.*, 1991; *Wagner*, 1992; *Poeter and Hill*, 1997] have been developed and are built into commercially available modeling software. Software packages such as PEST [*Doherty*, 2002] and UCODE [*Poeter et al.*, 2005] allow for automated model calibration that includes multiple datasets (e.g., hydraulic heads and tracer concentrations). Often, regression modeling requires that the inverse problem be overdetermined; hence only a handful of parameters can be estimated, or zonal patterns of heterogeneity need to be defined a priori. Stochastic inversion methods provide alternatives to conventional non-linear regression by seeking to identify multiple models that match a given dataset, thus yielding additional information on parameter uncertainty and how this translates into uncertainty in model predictions. Although papers on stochastic inversion abound in the hydrologic literature [e.g., *Ginn and Cushman*, 1990; *Harvey and Gorelick*, 1995; *McLaughlin and Townley*, 1996; *Gomez-Hernandez et al.*, 1997; *Capilla and Gomez-Hernandez*, 2003], widespread use of such methods has been hampered by the perceived complexity of these tools. In this volume, *Deutsch* provides an overview of common geostatistical approaches that were originally developed for petroleum and mineral problems but are applied with increasing frequency in subsurface hydrology. The author discusses practical aspects of geostatistical methods that range from estimation with sparse data and declustering, to integration of secondary data and complex geological structures. *Kitanidis* provides a review of a Bayesian framework for inversion of groundwater data,

Table 1. A partial list of information sources used for estimating hydrologic parameters or processes.

Data Type	Representative papers or books
Stratigraphic/sedimentologic information	*Weissmann and Fogg* [1999], *Koltermann and Gorelick* [1992]
Temperature	*Anderson* [2005], *Stonestrom and Constantz* [2003]
Geophysics	*Vereecken et al.* [2006], *Rubin and Hubbard* [2006]
Isotopes	*Clark and Fritz* [1997], *Kaufmann et al.* [1984]
Geochemistry and microbiology	*Chappelle* [2000], *Kendall and McDonnell* [1998]
Hydraulic head	*Hill and Tiedeman* [2007], *Kitanidis* [1997]
Solute concentrations	*Rubin* [2003], *Harvey and Gorelick* [1995]
Remote sensing	*Hoffmann et al.* [2003], *Houser et al.* [1998]

with emphasis on estimating hydraulic parameters using head data; this paper describes a linear Gaussian stochastic inverse approach (often referred to as geostatistical inversion) including the underlying concepts, mathematics, and applications.

The second section of this monograph includes papers that use data integration methods to characterize hydrologic properties such as hydraulic conductivity, porosity, or fracture connectivity as well as parameters representing boundary conditions and contaminant release histories. The interest in estimating hydrologic properties is many-fold, including development of models that can be used to assess the risks that contamination poses to potential receptors or to evaluate rates of natural processes including recharge. The papers collected here represent work with different data across a range of settings.

For vadose-zone applications, spatially variable water content controls flow in the subsurface. Extrapolation of these data to large spatial scales is complicated, however, given only direct measurements of water content. *Knight et al.* integrate neutron-probe and ground-penetrating radar data to assess specific geostatistical characteristics of water content data from Hanford, Washington, USA. This work moves toward more quantitative integration of surface GPR data in hydrologic studies, and offers insights into issues with the measurement support volume.

The hydrologic community has long benefited from shared interests and cross-pollination with petroleum engineering and exploration geophysics. This monograph includes two crossover papers from the petroleum community. *Avseth et al.* present a data integration method developed to characterize lithologic facies in reservoirs. Their approach combines geologic and seismic information using petrophysical relations within a Bayesian framework, while *Caers and Castro* present an application of a probabilistic approach to integrate geologic, facies, seismic, and well production data to characterization of a North Sea reservoir. To estimate geologic facies and match water and oil production data, they analyze static and dynamic data with multipoint geostatistical and perturbation methods. The work they present is applied to basin-scale fluid flow and reservoir dynamics; the methodologies, however, have direct application for hydrologic data integration. Multipoint geostatistics for data integration is still not commonly used in hydrology, despite work such as this that indicates its promise [e.g., *Feyen and Caers*, 2006].

Michalak and Shlomi contribute a theoretical framework for estimating the spatial and temporal evolution of solute plume distributions. This framework is based on geostatistical inverse modeling and multiple monitoring events, given knowledge about geological variability and other factors affecting solute transport, but without knowing the source location or release history. In their approach, concentration data can be used to reconstruct past plume distributions that are consistent with all available information. *Woodbury* presents the generalized inverse problem for heat and groundwater, as an example of how the Bayesian framework can be used for data integration. The paper includes two examples, one focusing on inversion of heat conduction for paleoclimate reconstructions, and the second focusing on groundwater flow within the Edwards aquifer.

In a vision paper, *Yeh et al.* discuss state-of-the-art tomographic approaches including both hydraulic tomography and electrical resistivity tomography. Several examples illustrate the benefits of combining multiple data types, such as hydraulic and tracer data. The authors then propose tomographic approaches to basin-scale hydrologic characterization; they suggest that natural hydrologic, geologic, and climatic stimuli (e.g., river-stage fluctuations, earthquakes, and lightning) can serve as hydrologic or geophysical perturbations needed for regional-scale tomographic surveys (i.e., hydraulic, seismic, or electrical).

In addition to characterizing physical or chemical properties that affect hydrologic processes, data integration methods are used to shed light on the processes themselves. The third section of the monograph is a collection of papers that demonstrate the use of diverse types of data to elucidate processes spanning subsurface-hydrologic research, from paleohydrology to watershed response to modern coastal aquifer dynamics. A range of data types (e.g., geochemical, isotopic, hydraulic, geophysical) and integration methods (i.e., spectral analysis, physically based numerical modeling, etc.) are considered. This range of topics is timely as we attempt to identify the influence of human activities associated with land use and climate change on hydrologic and ecological systems.

Hyndman et al. illustrate the use of the new Integrated Landscape Hydrology Model (ILHM), which was developed to predict spatial and temporal variations in groundwater recharge at the watershed scale. This code simulates the redistribution of precipitation through the vegetation canopy, sediment surface, soil and sediment layers, and snow pack to various surface and subsurface pathways using a process-based description of the water balance, based on GIS data and minimal use of site-specific parameters. A process-based simulation for a watershed in western Michigan, USA, illustrates the region's strong seasonality in recharge rates; most of the precipitation and snowmelt becomes groundwater recharge from September through March, while virtually none of the precipitation during the growing season is recharged.

The dynamics of coastal and island aquifers remain important basic- and applied-science topics. Understanding inter-

actions between aquifers, estuaries, and the coastal ocean requires consideration of many different data types collected over a range of temporal and spatial scales. Saltwater intrusion is a potential threat to many coastal and island aquifers, many of which are sole-source supplies of potable water. *Marksamer et al.* investigate the Nantucket Island aquifer in Massachusetts, USA, which extends deeper than expected given the current climate and water-table configuration. The authors use numerical modeling and multiple lines of evidence to test alternative paleohydrologic hypotheses to explain anomalous offshore freshwater and Nantucket's deep freshwater lens. Working in the coastal region of the southeastern USA, *Schultz et al.* combine groundwater monitoring, geochemical, electrical, electromagnetic, and vegetation mapping data to examine multi-scale, spatial and temporal coastal-aquifer dynamics. Target processes include saltwater intrusion, submarsh groundwater discharge, salinity gradients at the ocean boundary, and possible pore-water free convection.

The spectral content of hydrologic time series can provide insight into the time-scales of, and linkages between, important natural processes. *Kendall and Hyndman* demonstrate how spectral analysis of hydrologic datasets can be used to better understand linkages between precipitation, streamflows, and groundwater levels for watersheds in northern lower Michigan, USA. This analysis shows non-stationary behavior in these hydrologic systems, including the large reductions in summer streamflows due to canopy interception and evapotranspiration.

Fractured rock is, perhaps, the most complicated hydrologic setting [*National Research Council*, 1996]. Fluid concentration data from many fractured rock sites do not follow standard advective-dispersive behavior, and new data integration approaches are needed to identify dominant processes and understand the role of permeability heterogeneity [*National Research Council*, 2000, 2001b]. *Shapiro et al.* present an example of data integration from the U.S. Geological Survey's Fractured-Rock Hydrology research site, near Mirror Lake, New Hampshire, USA. The authors investigate anomalous solute-transport behavior at a variety of spatial scales using tracer and hydraulic testing as well as chemical sampling. Detailed borehole information and fracture mapping was integrated with the hydrologic data to clarify the geologic controls on flow and transport at each scale.

The collection of studies in this volume clearly demonstrates the value of data integration for hydrology; important limitations, however, remain. Recent work has underscored pitfalls and limitations of certain approaches or strategies used to combine data of different types. For example, additional work is needed to address the problems arising from model identification, non-linear feedbacks, uncertainty assessment, realistic characterization of geological variability, and discrepancies between the support volumes of different measurement types. The monograph's fourth section focuses on meta analysis and includes papers that reflect on opportunities for further research. *Singha et al.* discuss problems in the conversion of geophysical tomograms to hydrologic properties of interest. Although tomograms may provide qualitative information about hydrologic properties, the images have limited resolution and tend to be blurry versions of reality. The authors compare an analytical approach with a numerical approach to evaluate and address this problem. *McLaughlin* also discusses limitations associated with environmental data assimilation, in particular, problems that arise from the assumptions of linearity and normality on which most current approaches are based. He proposes that robust, rather than optimal, estimates should be sought, and that nonlinearity should be accepted and addressed.

Given current attention to coupled physical and chemical processes, and the increasing importance of groundwater as a resource, there is a strong need for novel data integration methods in hydrology. With continued advances in computational resources and rapidly evolving software for numerical modeling and inversion, future data integration methods will be better able to resolve both the nature of subsurface heterogeneities and the rates of critical processes across the range of hydrologic scales. Such developments will provide tools to help scientists address questions that arise through interdisciplinary research, where the measurements and models incorporate a host of processes that were typically studied individually within single disciplines. We believe that the integration of data and methods from hydrology with those from other sciences will be an active area of future research as hydrologic problems are increasingly recognized as being complex and dynamic. Data integration methods can provide important advances in the study of water quality and quantity, which will both be imperative for future decision-making in water resources.

REFERENCES

Anderson, M. P., Heat as a groundwater tracer, *Ground Water,* 6(43): 951–968, 2005.

Capilla, J. E. and J. J. Gomez-Hernandez, Stochastic inversion in hydrogeology, *Journal of Hydrology,* 281(4): 326 pp., 2003.

Chapelle, F. H., *Ground-water Microbiology and Geochemistry*, John Wiley and Sons, 468 pp., 2000.

Clark, I. D. and P. Fritz, *Environmental Isotopes in Hydrogeology*, CRC, 352 pp., 1997.

Doherty, J., *PEST—Model independent parameter estimation, version 6*, Queensland, Australia, Watermark Numerical Computing., 2002.

Feyen, L. and J. Caers, Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations, *Advances in Water Resources,* 29(6): 912–929, 2006.

Gailey, R. M., S. M. Gorelick and A. S. Crowe, Coupled process parameter estimation and prediction uncertainty using hydraulic head and concentration data, *Advances in Water Resources,* 14(5): 301–314, 1991.

Ginn, T. R. and J. H. Cushman, Inverse methods for subsurface flow; a critical review of stochastic techniques, *Stochastic Hydrology and Hydraulics,* 4(1): 1–26, 1990.

Gomez-Hernandez, J. J., A. Sahuquillo and J. E. Capilla, Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data; I, Theory, *Journal of Hydrology,* 203(1–4): 162–174, 1997.

Gorelick, S. M., Large scale nonlinear deterministic and stochastic optimization: Formulations involving simulation of subsurface contamination, *Mathematical Programming,* 48(1–3): 19–39, 1990.

Harvey, C. F. and S. M. Gorelick, Mapping hydraulic conductivity; sequential conditioning with measurements of solute arrival time, hydraulic head, and local conductivity, *Water Resources Research,* 31(7): 1615–1626, 1995.

Hill, M. C. and C. R. Tiedeman, *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, Wiley, 455 p., 2007.

Hoffmann, J., D. L. Galloway and H. A. Zebker, Inverse modeling of interbed storage parameters using land subsidence observations, Antelope Valley, California, *Water Resources Research,* 39(2): 1031, doi:10.1029/2001WRR001252, 2003.

Houser, P. R., W. J. Shuttleworth, J. S. Famiglietti, H. V. Gupta, K. H. Syed and D. C. Goodrich, Integration of soil moisture remote sensing and hydrologic modeling using data assimilation, *Water Resources Research,* 34(12): 3405–3420, 1998.

Hubbard, S. and G. Hornberger, Introduction to special section on Hydrologic Synthesis, *Water Resources Research,* 42: W03S01, doi:10.1029/2005WR004815, 2006.

Kaufmann, R., A. Long, H. Bentley and S. Davis, Natural chlorine isotope variations, *Nature,* 309: 338–340, 1984.

Kendall, C. and J. J. McDonnell, *Isotope tracers in catchment hydrology*, Elsevier, 1998.

Kitanidis, P. K., *Introduction to geostatistics; applications to hydrogeology*, Cambridge, Cambridge University Press, 249, 1997.

Koltermann, C. and S. M. Gorelick, Paleoclimatic signature in terrestrial flood deposits, *Science,* 256: 1775–1782, 1992.

McLaughlin, D. and L. R. Townley, A reassessment of the groundwater inverse problem, *Water Resources Research,* 32(5): 1131–1161, 1996.

National Research Council, *Rock Fractures and Fluid Flow: Contemporary Understanding and Applications*, Washington D.C., National Academy Press, 551 pp., 1996.

National Research Council, *Seeing into the Earth: Noninvasive Characterization of the Shallow Subsurface for Environmental and Engineering Application*, Washington D.C., National Academy Press, 129 pp., 2000.

National Research Council, *Basic Research Opportunities in Earth Science*, Washington D.C., National Academy Press, 168 pp., 2001a.

National Research Council, *Grand Challenges in Environmental Sciences*, Washington D.C., National Academy Press, 2001b.

Neuman, S. P. and S. Yakowitz, A Statistical Approach to the Inverse Problem of Aquifer Hydrology, 1. Theory *Water Resources Research,* 15(4): 845–860, 1979.

Poeter, E. P. and M. C. Hill, Inverse models; a necessary next step in ground-water modeling, *Ground Water,* 35(2): 250–269, 1997.

Poeter, E. P., M. C. Hill, E. R. Banta, S. Mehl and S. Christensen, *UCODE_2005 and Six Other Computer Codes for Universal Sensitivity Analysis, Calibration, and Uncertainty Evaluation*, U.S. Geological Survey Techniques and Methods. 6-A11: 283 pp., 2005.

Rubin, Y., *Applied Stochastic Hydrogeology*, Oxford University Press, 416 pp., 2003.

Rubin, Y. and S. S. Hubbard, Eds., *Hydrogeophysics (Water and Science Technology Library)*. Netherlands, Springer, 2006.

Stonestrom, D. A. and J. Constantz, *Heat as a tool for studying the movement of ground water near streams*, USGS Circular 1260: 105 pp., 2003.

Theis, C. V., The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using ground-water storage, *American Geophysical Union Transcript,* 16: 519–524, 1935.

Vereecken, H., A. Binley, G. Cassiani, A. Revil and K. Titov, Eds., *Applied Hydrogeophysics*, NATO Science Series: IV: Earth and Environmental Sciences, Springer-Verlag, 2006.

Wagner, B. J., Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling, *Journal of Hydrology,* 135: 275–303, 1992.

Weissmann, G. S. and G. E. Fogg, Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework, *Journal of Hydrology,* 226(1–2): 48–65, 1999.

A Review of Geostatistical Approaches to Data Fusion

Clayton V. Deutsch

University of Alberta, Edmonton, Alberta, Canada

Geostatistics has evolved to a mature discipline with a well understood theoretical framework and a standard set of tools. The tools have been applied with many geospatial variables in many different contexts. This paper provides a brief review of geostatistical approaches to problems involving multiple data types in subsurface hydrology. The random function paradigm of geostatistics is presented. Bayes Law is the engine that permits multivariate spatial and remotely sensed data to be integrated. The required multivariate probabilities are often fit with the Gaussian distribution. There are many implementation decisions and practicalities of geostatistics. These include declustering, inference in presence of sparse data, dealing with many secondary data, and modeling complex geological features. Subjects of practical importance are reviewed.

1. INTRODUCTION

The word *geostatistics* commonly refers to the theory of regionalized variables and the related techniques that are used to predict rock properties at unsampled locations. Georges Matheron formalized this theory in the early 1960's (Matheron, 1971). The development of geostatistics was led by engineers and geologists faced with real problems. They were searching for a consistent set of numerical tools that would help them with ore reserve estimation, reservoir performance forecasting, and site characterization.

At any instance in geological time, there is a single true distribution of rock properties over each study area. This true distribution is inaccessible with limited data and the chaotic nature of certain aspects of geological processes. Geostatistics strives to create numerical models that mimic the physically significant features of property variations.

Conventional mapping algorithms were devised to create smooth maps to reveal large-scale geologic trends; they are low pass filters that remove high frequency property variations. For practical problems of flow prediction, however, this variability has a large affect on the predicted response. Geostatistical simulation techniques, conversely, were devised with the goal to reproduce a realistic amount of variability, that is, create maps or realizations that are neither unique nor smooth. Although the small-scale variability of these realizations may mask large-scale trends, geostatistical simulation is more appropriate for predictions of subsurface flow.

Geostatistics is primarily concerned with constructing high-resolution 3-D models of categorical variables such as facies and continuous variables such as porosity and permeability. It is necessary to have *hard* truth measurements at some volumetric scale. All other data types including geophysical data are called *soft* data and must be calibrated to the hard data. It is neither possible nor optimal to construct models at the resolution of the hard data. Models are generated at some intermediate geological modeling scale, and then scaled to an even coarser resolution for flow modeling. An important goal of geostatistics is the creation of detailed numerical 3-D geologic models that simultaneously account for a wide range of relevant data of varying degrees of resolution, quality, and certainty. Much of geostatistics relates to data calibration and reconciling data types at different scales. This data integration or *fusion* is the focus of this review paper.

Subsurface Hydrology: Data Integration for Properties and Processes
Geophysical Monograph Series 171

10.1029/171GM03

Geostatistical techniques allow alternative realizations to be generated. These realizations are often combined in a model of uncertainty, that is, they are processed through a numerical model of the response and the different outcomes are assembled in a distribution of response uncertainty. Uncertainty is becoming an important goal of geostatistical studies.

Numerical models are rarely built in one step. A hierarchical framework is followed with different techniques and tools at each level. A typical scenario consists of (1) mapping large scale bounding surfaces with conventional or geostatistical techniques, (2) mapping trends of facies proportions within each major stratigraphic layer, (3) creating high resolution facies models within each layer reproducing the mapped trends, (4) assigning continuous rock properties such as porosity and permeability within each facies, and (5) post processing and upscaling the resulting high resolution models for flow simulation. The classical random function model formalism of geostatistics is presented first, then some of the practical implementation aspects are described.

2. RANDOM FUNCTION FORMALISM

We start by considering a regionalized variable such as a subsurface elevation, formation thickness, facies proportion, facies indicator, porosity or permeability. We denote a specific value as z. The uncertainty about an unsampled value z is modeled through the probability distribution of a random variable (RV) Z. The probability distribution of Z after data conditioning is usually location-dependent; hence the notation $Z(\mathbf{u})$, with $\mathbf{u}$ being the coordinate location vector. A random function (RF) is a set of RVs defined over some field of interest, e.g., $Z(\mathbf{u})$, $\mathbf{u} \in$ study area A. Geostatistics is concerned with inference of statistics related to a random function (RF).

Inference of any statistic requires some repetitive sampling. For example, repetitive sampling of the variable $z(\mathbf{u})$ is needed to evaluate the cumulative distribution function: $F(\mathbf{u};z) = \text{Prob}\{ Z(\mathbf{u}) \leq z \}$ from experimental proportions. However, in most cases, at most one sample is available at any single location $\mathbf{u}$; therefore, the paradigm underlying statistical inference processes is to trade the unavailable replication at location $\mathbf{u}$ for replication over the sampling distribution of z-samples collected at other locations within the same general area.

This trade of replication corresponds to the decision of stationarity. Stationarity is a property of the RF model, not of the underlying regionalized variable. Thus, it cannot be checked from data. The decision to pool data into statistics across facies is not refutable a priori from data; however, it can be shown inappropriate a posteriori if differentiation per facies is critical to the study.

The first and most important aspect of stationarity is the decision to pool data together for common processing. Another aspect of stationarity is a decision regarding the location-dependency of statistical parameters. A common practical approach is to assume that key statistical parameters do not depend on location within reasonably defined geological populations.

The statistical paradigm faced by geostatisticians is one of multivariate statistics: the same variable at multiple locations and multiple secondary data. We could denote the secondary data as $Y(\mathbf{u})$ and index Y if required to be clear regarding the number of secondary data. This is illustrated schematically in Plate 1.

The two wells and gridded seismic response on Plate 1 illustrate the multivariate aspect of the problem faced by geostatisticians. We are interested in the uncertainty at a location that has not been drilled. The nearby data (n) consist of well and seismic data:

$$(n) = \{z(\mathbf{u}_\alpha), \alpha=1,\ldots,n_w\}, \{y(\mathbf{u}_\beta'), \beta=1,\ldots,n_s\} \qquad (1)$$

The uncertainty at a particular unsampled location must be inferred in light of the (n) conditioning data. A best estimate can be retrieved from the conditional distribution or it could be sampled by Monte Carlo simulation for alternative realizations. The standard approach to estimate conditional probabilities is Bayes Law, which has been used for more than 200 years. Bayes Law provides the arithmetic to infer the conditional distribution of the unsampled value $z(\mathbf{u})$:

$$F_{Z(\mathbf{u})|(n)}(z) = \frac{F_{Z(\mathbf{u}),(n)}(z_0, z_1, \ldots, z_n)}{F_{(n)}(z_1, \ldots, z_n)} \qquad (2)$$

The numerator on the right side is an $n+1$ variate distribution of the unknown and the n data. The denominator on the right side is the n variate distribution of the conditioning data. The univariate distribution on the left side is what we are after–the conditional distribution of the unsampled value given the set of conditioning data (n).

Inference of the required multivariate distributions is virtually impossible. There are no replications of the unsampled value with the data values and there are unlikely to be replications of the precise data configuration (n). Nevertheless, those multivariate probabilities are required for inference of the conditional distribution.

The required multivariate probabilities are calculated from either an analytical distribution model or from a large set of analogue data deemed representative (sometimes referred to as a training image). The conventional paradigm of geostatistics is to use analytical distributions with parameters inferred from the available data. The multivariate Gaussian distribution will

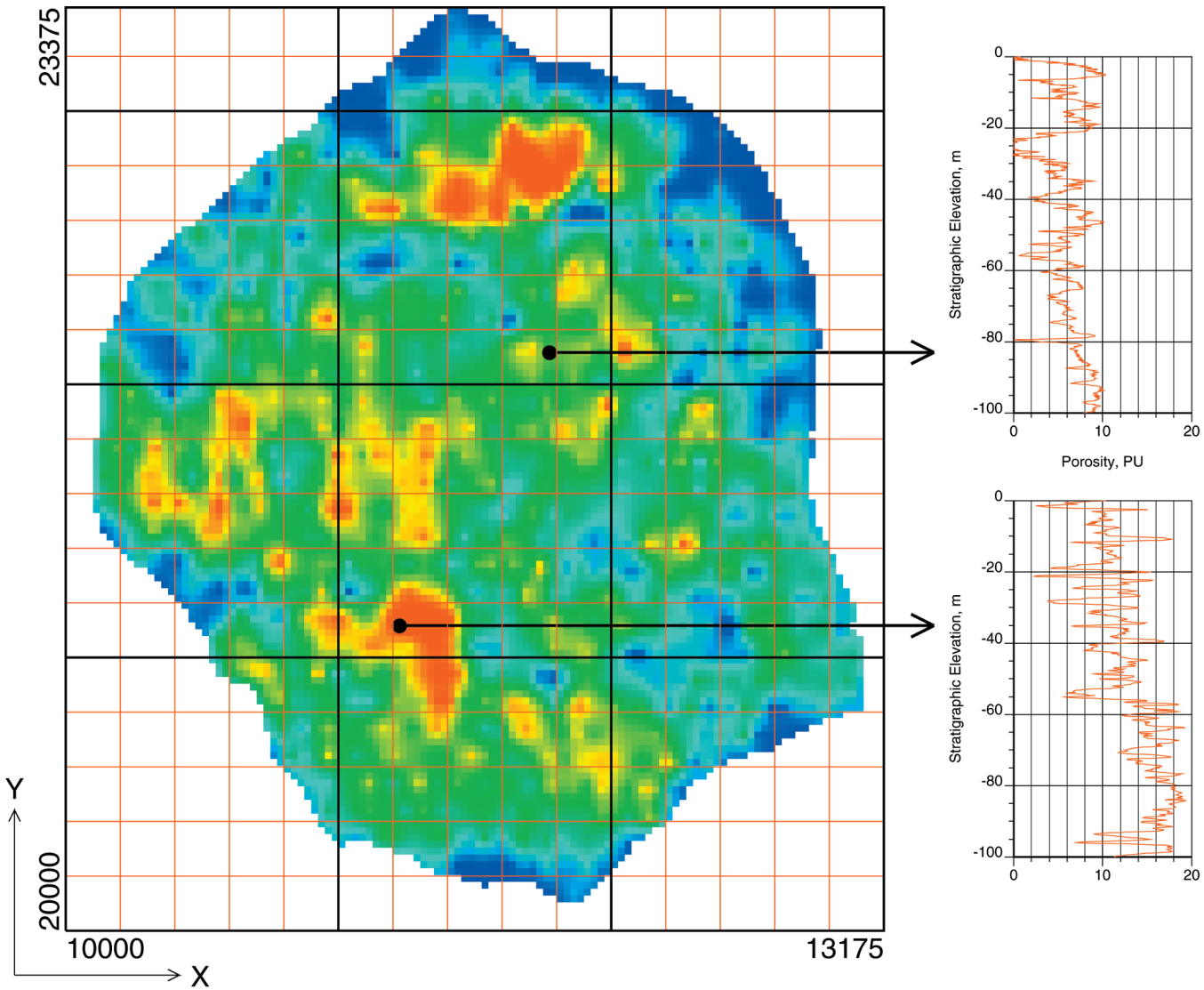


Plate 1. Illustration of the typical case faced by geostatisticians: there are a limited number of locations with precise measurements (the two wells in this case) and secondary variables that are often on grids (one variable shown).

be explained in the next section. The classical approach of variograms and kriging will be explained now.

Subsurface variables are heterogeneous. Their spatial variability is quantified by the variogram function:

$$2\gamma(\mathbf{h}) = E\left\{\left[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right]^2\right\} \quad (3)$$

2γ is variability and is in the units of variance, $\mathbf{h}$ is a vector distance, $Z(\mathbf{u})$ is the random variable. The expected value is approximated by a discrete sum over the available pairs. The available data do not permit estimation of 2γ for many distance and direction lags. The function is fit to interpolate 2γ for $\mathbf{h}$-values that cannot be calculated.

Estimation can be formulated as an optimization problem. The linear estimate at unsampled location $\mathbf{u}_0$ is written:

$$\left[z^*(\mathbf{u}_0) - m(\mathbf{u}_0)\right] = \sum_{i=1}^{n} \lambda_i \cdot \left[z(\mathbf{u}_i) - m(\mathbf{u}_i)\right] \quad (4)$$

$m(\mathbf{u})$ is the location-dependent mean and the λs are weights that are calculated to minimize the expected error variance. The equations that lead to the optimal weights are referred to as the normal equations or the simple kriging equations:

$$\sum_{j=1}^{n} \lambda_j \cdot C(\mathbf{u}_i - \mathbf{u}_j) = C(\mathbf{u}_i - \mathbf{u}_0), \; i = 1,...,n \quad (5)$$

The $C(\mathbf{h})$ covariance values are derived from the variogram through the relation $C(\mathbf{h})=\sigma^2-\gamma(\mathbf{h})$, which is valid with the assumption of stationarity. Relatively straightforward modifications are necessary if the decision of stationarity is relaxed. Constraints may be added to ensure unbiasedness without specifying the location-dependent mean; the modifications are straightforward. We can calculate the minimized error variance, but it has no practical meaning outside of the Gaussian context (see Section 3 below).

Estimates from Equation 4 are useful. They provide a useful means to construct a grid of estimates. These estimates are used for resource assessment and visualization of geologic trends. The kriging formalism of Equations 4 and 5 may be extended to multiple correlated variables. The result is cokriging. The estimate follows the same form; however, the covariance values must come from a mathematically valid model of coregionalization.

Kriging was state of the art in the late 1970s and early 1980s. Geostatisticians have come to expect more from their numerical models: local and joint uncertainty. Equation 2 is valid. A multivariate distribution is required. The multivariate Gaussian distribution is a remarkably tractable model that has come to be relied upon in geostatistical calculations.

3. MULTIVARIATE GAUSSIAN DISTRIBUTION

The multivariate probabilities required for inference of continuous variable uncertainty cannot be directly inferred from data. A multivariate Gaussian model is systematically adopted. The continuous variable is transformed to a Gaussian distribution, and then all multivariate distributions are assumed to be Gaussian. We would wish for alternative probabilistic models to choose from; however, the multivariate Gaussian probability distribution is remarkably tractable and used almost exclusively. Figure 1 illustrates transformation of a continuous variable from an arbitrary distribution to a Gaussian distribution. The distributions are shown as cumulative distribu-

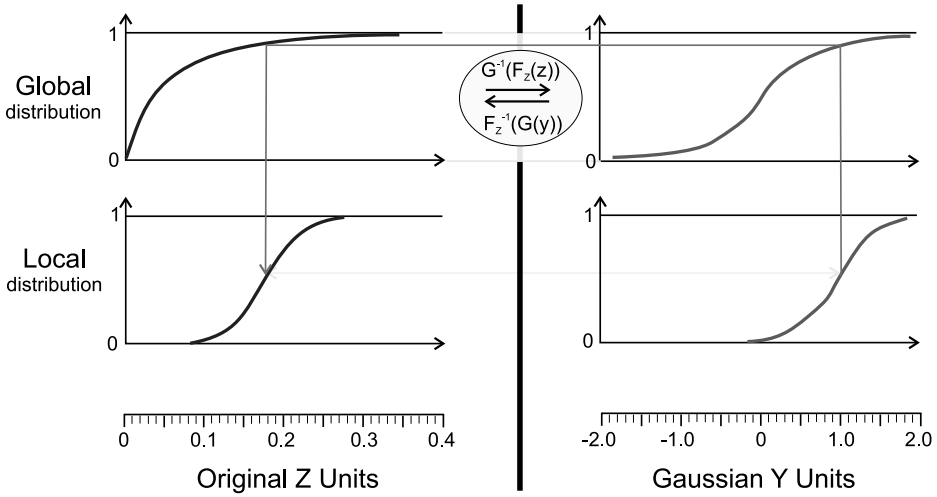


Figure 1. Schematic illustration of normal score transform. The original Z- data are on the left and the Gaussian Y-values are on the right. The top figures are the global CDFs and the bottom figures represent local CDFs. Quantiles are transformed using the global distribution (the three part blue line).

tions. In Gaussian units (the right side), all distributions are Gaussian in shape. The uncertainty in original units must be established by back transformation. The transformation and back transformation are written as:

$$y = G^{-1}(F(z)) \text{ and } z = F^{-1}(G(y)) \quad (6)$$

Figure 1 reveals an important point. All conditional distributions in Gaussian units are non-standard Gaussian, see the lower right. The quantiles of such distributions can be back transformed via the global transformation. Conditional distributions in original units are not Gaussian, but we can establish their shape numerically, that is, back transformation of many quantiles. The 99 percentiles would be a good start; more are required for a stable estimate of the variance.

The mean and variance of each conditional non-standard Gaussian distribution are calculated with the normal equations that are identical to the kriging equations given in equations 4 and 5. The stationary mean is set to 0.0 in Gaussian units and the variogram/covariance are calculated from the normal score transforms of the data. The variance of estimation has particular meaning in the Gaussian case; it is the variance of the conditional distribution:

$$\sigma^2 = 1 - \sum_{i=1}^{n} \lambda_i \cdot C\left(u_i - u_0\right) \quad (7)$$

A small example will be developed at the expense of some space. This example is a classic illustration of modern geostatistical tools used to assess uncertainty.

3.1 Small Example

Consider a square grid of 101–16m grid cells that cover just over one regular Section of land. Let's directly model porosity. The global representative distribution will be taken as lognormal with a mean $m=0.15$ and a standard deviation $\sigma=0.075$. The global representative distribution would be obtained by declustering and/or debiasing using the available well and seismic data. Consider an average data of 0.15 in the northwest corner of the area and a high data of 0.25 in the southeast corner of the area.

Uncertainty is characterized in Gaussian units. The transformation to a standard Gaussian distribution is defined analytically in this case:

$$y = \frac{\log(z) - \alpha}{\beta}$$
$$\beta = \sqrt{\log\left(1 + \frac{\sigma^2}{m^2}\right)} \text{ and } \alpha = \log(m) - \beta^2/2 \quad (8)$$

In our case $\alpha = -2.01$ and $\beta = 0.472$. The back transform is also defined analytically: $z = exp(y\beta + \alpha)$. The porosity data values of 0.15 and 0.25 are transformed to 0.236 and 1.317, respectively.

A fitted variogram model of the Gaussian transformed values is required. This would be obtained from the available data and analogue information. The variogram will be taken as an exponential function with an effective range of 2000m: $\gamma(\mathbf{h}) = 1 - exp(3\mathrm{h}/2000)$. In fact, $\gamma(\mathbf{h})$ is the semivariogram or one half of the variogram. Under a decision of stationarity, the covariance function is $C(\mathbf{h}) = 1 - \gamma(\mathbf{h}) = exp(3\mathrm{h}/2000)$.

Local conditional distributions are defined everywhere by a local conditional mean and variance that are computed by simple kriging. Plate 2 shows these results. The locations of the wells are evident on the conditional variance map–the conditional variance is zero at the two well locations. These results are in Gaussian units. We back transform these conditional distributions to original units by back transforming a large number of quantiles, say 200. Plate 3 shows maps of the conditional mean, conditional variance, P_{90} low value and P_{10} high value in original units. Note how the conditional variance in original units is higher in the south and east because the mean is higher; the conditional variance in original units depends on the data as well as the data configuration.

Simulated realizations are required for two reasons. Firstly, they provide numerical models of heterogeneity for process evaluation. Secondly, they permit input uncertainty to be transferred to output uncertainty, for example, calculating uncertainty in resources or transport. There are a number of implementations that generate multiple realizations. Sequential methods such as sequential Gaussian simulation are popular.

Multiple realizations of porosity are generated by Gaussian simulation. Five realizations are shown on the left of Plate 4. The two well data are reproduced by all realizations. The pore volume was calculated on each realization assuming a thickness of 10m. The distribution of pore volume is shown at the right of Plate 4. These realizations allow us to visualize heterogeneity as well as assess uncertainty. The realizations could be ranked by their pore volume and select realizations (say the ones with the P_{90}, P_{50} and P_{10} outcomes) could be input to flow simulation.

This little example shows a hint of what geostatistics is aimed at. In practice, we must consider multiple stratigraphic layers, multiple facies, multiple data types, and multiple variables such as residual saturation and permeability. Some practicalities are addressed below.

3.2 Block Cokriging

An important practical reality of geostatistics is the presence of data with different type, noise content, and volume

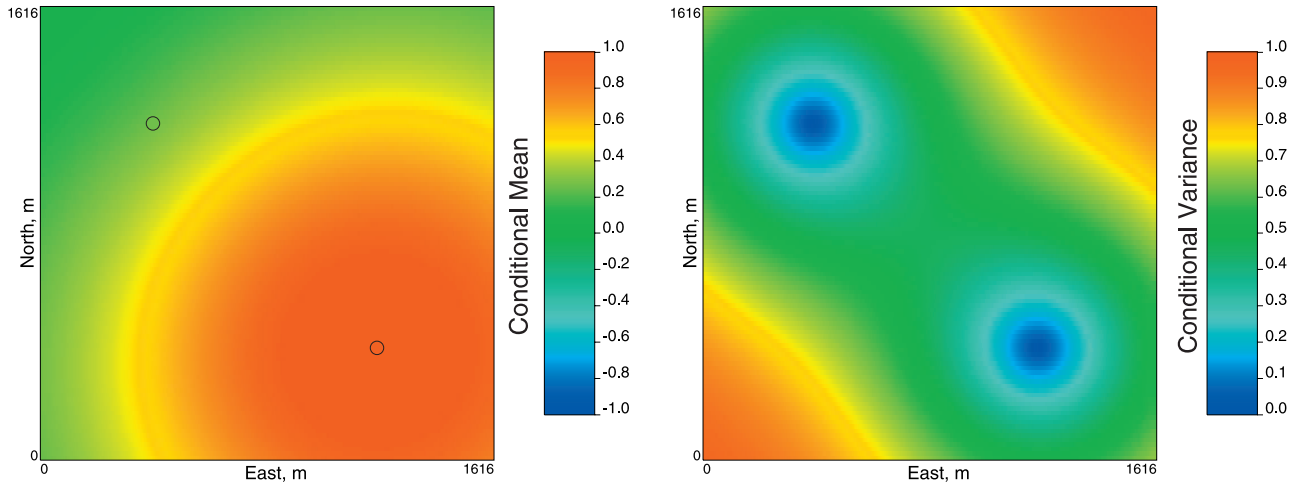


Plate 2. Map of the conditional mean (left side) and conditional variance (right side) for the Small Example.

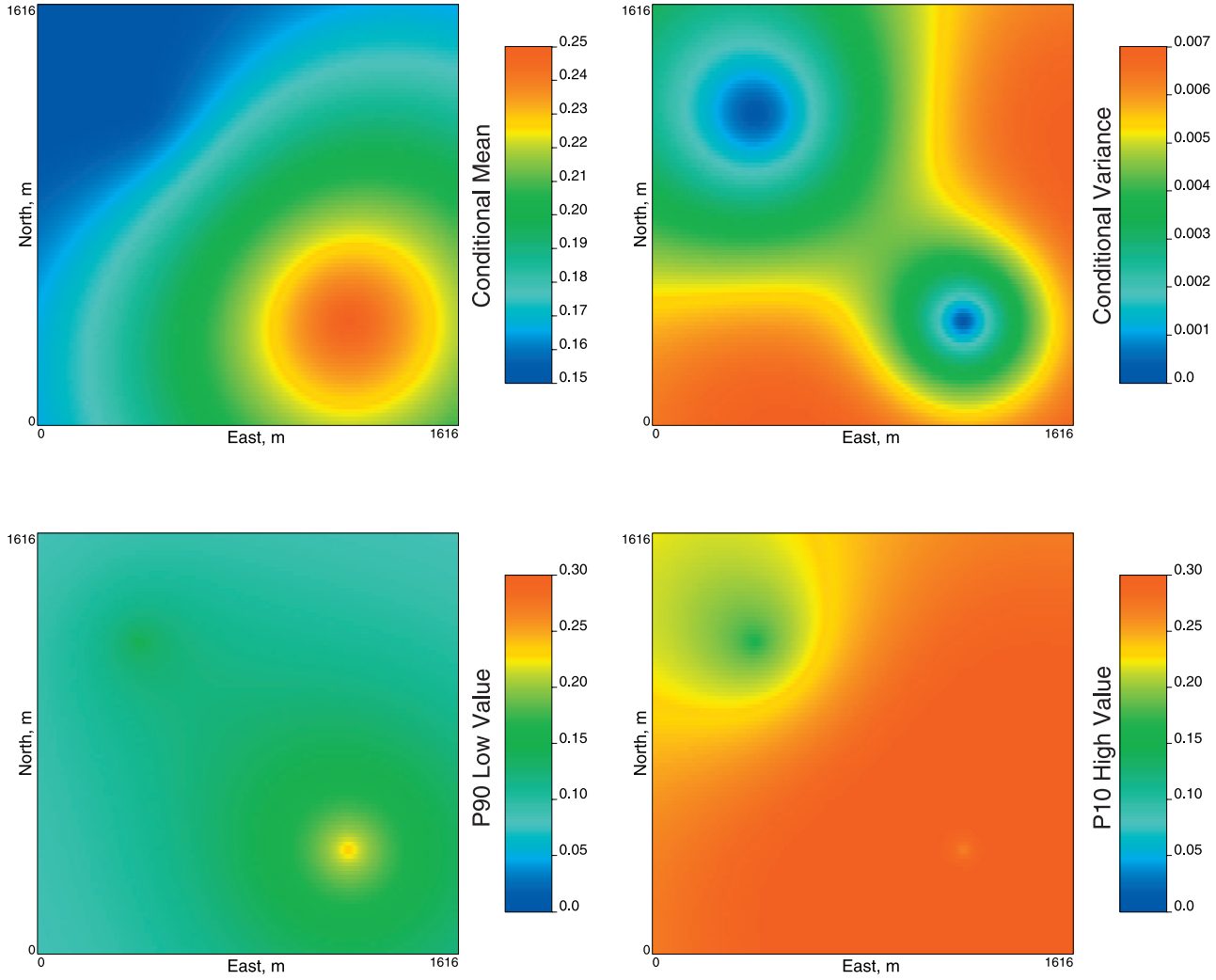


Plate 3. Map of the conditional mean (upper left) and conditional variance (upper right) in original units. Maps of the P90 low value and P10 high values are shown in the lower left and right.

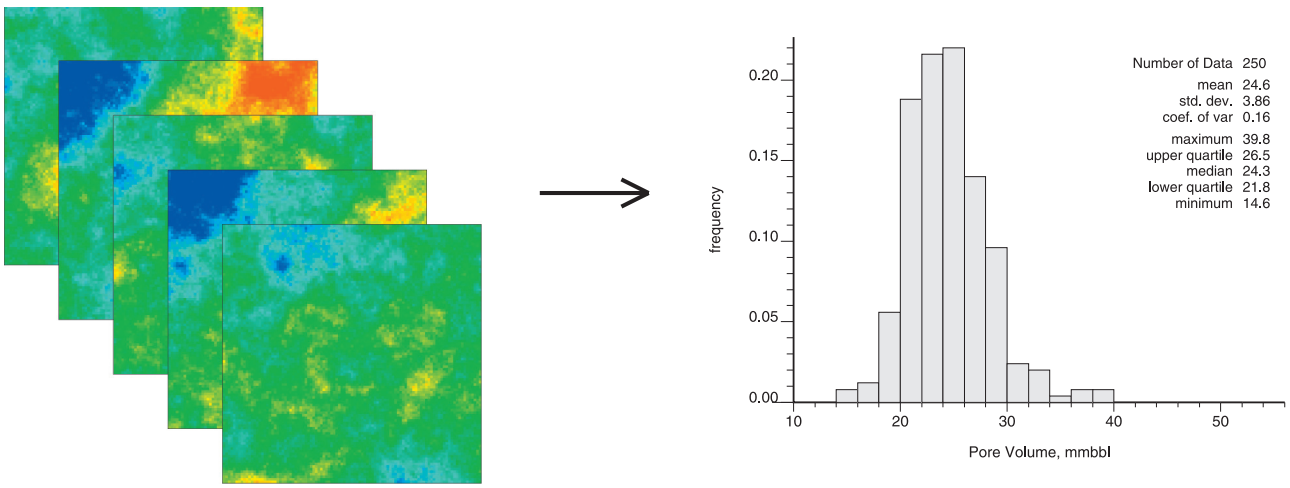


Plate 4. Multiple realizations (5 out of 250) are illustrated on the left and a histogram of the OOIP for the 250 realizations is shown to the right.

scale. We could invoke block cokriging to address these three critical issues. Different data types are handled with a cokriging and a model of coregionalization. Different volume scales of measurement are handled by block cokriging, that is, the use of volume averaged covariances. There are a number of inference problems and challenges with this approach: (1) linear averaging is assumed in Gaussian units, which is only correct if the original variable histograms are Gaussian in shape, (2) the point-scale statistics including histograms and variograms must be known, and (3) the noise content of each data source must also be known. This approach is valid and manageable in many cases. Nevertheless, these assumptions are serious and often lead practitioners to consider some simplifications. A number of practical implementation issues will now be discussed. These are unquestionably important for reasonable results in the combination of data with geostatistics.

4. PRACTICAL IMPLEMENTATION

4.1 Representative Statistics

Wells are not drilled to be statistically representative of the site; they are often intended as locations for production. Even in preliminary appraisal, there is a desire to delineate interesting areas of the site. It is critical to establish a representative distribution for each variable being modeled. This includes facies proportions and the histograms of porosity and permeability within each facies type. *Declustering* techniques weight the data such that wells drilled close together are given less weight. Wells drilled farther apart are given more weight. Declustering is suitable when there are sufficient data to sample areas of high and low quality. Sometimes there are too few wells. There may be areas of relatively poor reservoir quality that have not been drilled. *Debiasing* techniques are used to establish representative distributions based on a secondary variable such as seismic or a geologic trend. The results of declustering and debiasing include representative facies proportions and representative histograms of each continuous variable under consideration. A large-scale trend model may have been built for debiasing–this trend model will also come into subsequent geostatistical calculations.

An essential feature of geostatistics is inference in presence of sparse data. We are faced with a paradox. A lack of data is precisely when a geostatistical model of uncertainty is warranted; however, it is also the case when inferring required parameters is difficult. Limiting ourselves to statistics we can infer from the available data would be a mistake. We must often use analogue information related to spatial continuity, particularly in the vast interwell region. The spatial continuity in the vertical direction is relatively easy to infer even with limited well data. Horizontal to vertical anisotropy ratios based on the geologic setting can be useful to infer the horizontal continuity. The vertical variogram shape is used, but scaled according to a ratio. Figure 2 shows some typical ratios (Deutsch, 2003).

4.2 Hierarchical Modeling

A sequential approach is often followed for reservoir modeling. The large-scale features are modeled first followed by smaller, more uncertain, features:

(1) Establish the stratigraphic layers to model, that is, define the geometry of the *container* being modeled. A conceptual model for the large scale continuity of facies and petrophysical properties within each major layer is chosen.
(2) The bounding surfaces are mapped. They may be simulated with geostatistical techniques if they are associated with considerable uncertainty.
(3) The facies rock types are modeled by cell-based or object-based techniques within each stratigraphic layer (see below). Multiple realizations represent uncertainty in facies.
(4) The porosity and other petrophysical variables are modeled on a by-facies basis. These may be modeled one after another or all together. Multiple realizations are used to represent uncertainty.
(5) The models are revised to match dynamic data such as pumping tests and flow history. Knowledge gained from trying to match this data may be coded as spatial constraints and the modeling repeated.
(6) These set of multiple realizations are input to flow and transport modeling or simply visualized to aid in decision making and resource assessment.

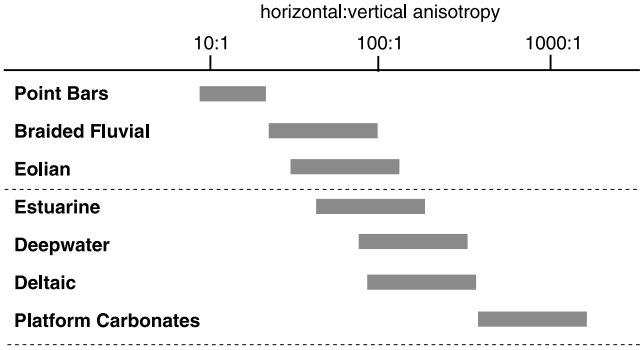


Figure 2. Some typical horizontal-to-vertical anisotropy ratio conceptualized from available literature and experience. Such generalizations can be used to verify actual calculations and supplement very sparse data.

A detailed description of these steps is beyond the scope of this review paper, but some of the references are suitable. The book by Chiles and Delfiner (1999) is a comprehensive overview of geostatistical techniques. The book by Cressie (1991) presents a statistical perspective on this approach. The book by David (1977) is a practical mining approach. The two books by Deutsch (1998 and 2003) provide a software and petroleum perspective, respectively. Goovaerts (1997) provides another comprehensive overview of geostatistical techniques. Isaaks and Srivastava (1989) provide a nice introduction to basic concepts. Journel and Huijbregts (1978) provide a comprehensive theoretical presentation from a mining perspective. Kitanidis (2000) provides an introduction from a hydrogeologic perspective.

4.3 Facies Modeling

Facies are often important in reservoir modeling because the petrophysical properties of interest are highly correlated with facies type. Facies are distinguished by different grain size or different diagenetic alteration. The facies must have a significant control on the porosity and other properties of interest; otherwise, modeling the 3-D distribution of facies will be of little benefit since uncertainty will not be reduced and the resulting models will have no more predictive power. An additional constraint on the choice of facies is that they must have straightforward spatial variation patterns. The distribution of facies should be at least as easy to model as the direct prediction of petrophysical properties. Once the facies are defined, relevant data must be assembled and a 3-D modeling technique selected.

The alternatives are (1) cell-based geostatistical modeling, (2) object-based stochastic modeling, or (3) deterministic mapping. Deterministic mapping is always preferred when there is sufficient evidence of the facies distribution to remove any doubt of the 3-D distribution. In many cases, there is evidence of geologic trends, which should be included in stochastic facies modeling.

Cell-based techniques are commonly applied to create facies models. The popularity of cell-based techniques is understandable: (1) local data are reproduced by construction, (2) the required statistical controls (variograms) may be inferred from limited well data, (3) soft seismic data and large-scale geological trends are handled straightforwardly, and (4) the results appear realistic for geological settings where there are no clear geologic facies geometries, that is, when the facies are diagenetically controlled or where the original depositional facies have complex variation patterns. Of course, when the facies appear to follow clear geometric patterns, such as sand-filled abandoned channels or lithified dunes, object-based facies algorithms should be considered.

From a geological perspective, it is convenient to view reservoirs and aquifers from a chrono-stratigraphic perspective. The sedimentary architecture is considered in light of a hierarchical classification scheme. We consider modeling this genetic hierarchy of heterogeneities by surfaces and objects representing facies associations.

Despite the realism of object-based modeling, many reservoirs show very complicated architectural element configurations developed during meander migration punctuated by avulsion events. It is becoming increasingly common to attempt facies modeling in a manner that mimics original deposition and alteration. Like object-based modeling, there is a perception that these process-based models are difficult to condition to well data.

Image analysis based techniques using multiple point statistics have evolved to use the models generated by object- and process-based models as training images. The features of such models are imposed on 3-D geocellular models with multiple point statistics (Guardiano and Srivastava, 1992).

4.4 Secondary Data

The block cokriging approach mentioned above has limited applicability in presence of many secondary data at different scales. Inference of the required statistics is virtually impossible. Collocated cokriging simplifies the process to consider collocated secondary variables; however, there is no simple way to consider a large number of secondary variables simultaneously.

Many different variables must be considered: small scale well data, large-scale remotely sensed variables, interpreted trend-like variables, and other response variables. These data often cover different areas, provide data at different scales, and are variably correlated together. Conventional geostatistical techniques, such as the block cokriging mentioned above, incorporate the spatial structure but these techniques are cumbersome in the presence of many secondary variables. An increasingly common approach is to merge all secondary data into a single variable that contains all of the secondary variable information; this provides a conditional distribution. The spatial distribution of each variable is mapped with data of the same type of information; this provides a second conditional distribution. The two conditional distributions are merged to provide updated posterior distributions. This merging is done in Gaussian units and the variables must be back transformed for final analysis.

Two Gaussian conditional distributions may be merged to an updated Gaussian distribution assuming conditional independence of the two distributions. This type of Markov model is very common. The parameters of the updated Gaussian distribution are given by:

$$m_U = \frac{m_1 \bullet \sigma_2^2 + m_2 \bullet \sigma_1^2}{\sigma_1^2 - \sigma_1^2\sigma_2^2 + \sigma_2^2} \text{ and } \sigma_U^2 = \frac{\sigma_1^2 \bullet \sigma_2^2}{\sigma_1^2 - \sigma_1^2\sigma_2^2 + \sigma_2^2} \quad (9)$$

This simple result is at the heart of much data integration. An important notion of data integration is that corroborating data cause the updated distributions to be non-convex. Figure 3 shows three examples. In the first case, both distributions are high (m_1=0.8, σ^2_1=0.6, m_2=1.0, σ^2_2=0.4), that is, the mean values are greater than the global mean; therefore, the updated distribution is quite high (m_u=1.21, σ^s_u=0.316). In the second case, one distribution is high and the other low (m_1=-0.5, σ^2_1=0.6, m_2=0.5, σ^2_2=0.6); therefore, the result is in the middle (m_u=0.0, σ^s_u=0.429). In the third case, both distributions are low (m_1=-0.8, σ^2_1=0.6, m_2=-1.0, σ^2_2=0.4); therefore, the updated distribution is even lower (m_u=-1.21, σ^s_u=0.316).

Recall that all of our data are transformed to Gaussian units, uncertainty is assessed, and the resulting uncertainty is back transformed to original units. Quantiles of local distributions can be back transformed or entire simulated realizations are back transformed. The multivariate Gaussian distribution is used routinely in geostatistics because it is straightforward to infer the required parameters with few data. Data integration and uncertainty prediction is relatively easy. Moreover, it is common that data within reasonably defined facies are often Gaussian. Nevertheless, we often seek an alternative to the Gaussian model to handle more complex features. The most common alternative is the indicator formalism (Journel, 1983).

4.5 Indicator Formalism

Indicators are applied to both continuous and categorical variables. A series of threshold values z_c are used to discretize the range of variability of the continuous Z-variable. The indicator coding of continuous variables:

$$i(\mathbf{u}; z_c) = \begin{cases} 1, \text{if } z(\mathbf{u}) \le z_c \\ 0, \text{otherwise} \end{cases} \quad (10)$$

This amounts to coding the continuous data as a series of cumulative probability values. It is common to consider between 9 and 20 threshold values; less than 9 leads to poor resolution and greater than 20 leads to difficult inference of the required parameters and no significant increase in precision of calculated conditional distributions.

Variogram analysis is conducted for each threshold. This permits the continuity of the low and high values to be modeled differently. The variograms should be consistent since they are based on the same underlying continuous variable; however, they are more flexible than the simplistic Gaussian model.

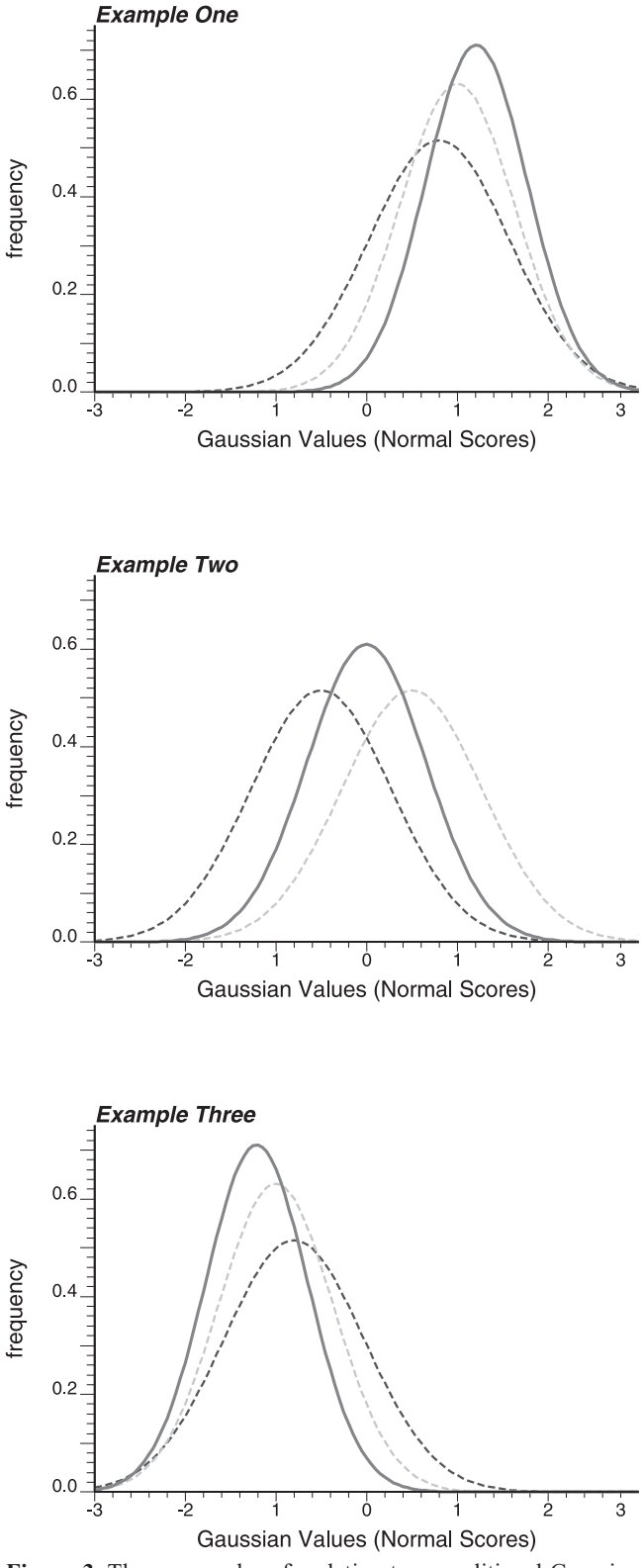


Figure 3. Three examples of updating two conditional Gaussian distributions into updated distributions.

Kriging is applied at each threshold with the corresponding indicator variogram to directly calculate an estimate of the CDF value at the threshold values. This leads to a direct estimate of the conditional distribution. Figure 4 shows a schematic example. It is necessary to ensure that the estimated CDF values form a valid distribution (non decreasing between 0 and 1) and to interpolate and extrapolate the CDF beyond the values predicted at the thresholds. These distributions can be used directly for uncertainty assessment or used in simulation to assess joint uncertainty.

The indicator coding for categorical variables is similar. Consider K facies. The data are coded as the probability of occurrence:

$$i(\mathbf{u};k)=\begin{cases}1, \text{if facies } k \text{ at location}\\ 0, \text{otherwise}\end{cases} \tag{11}$$

As with continuous variables, variograms are constructed for each of the K indicators. Kriging can be applied to predict the probability of each facies at an unsampled location. The probability estimates are corrected if necessary to ensure that they are non negative and sum to one. They are then used for uncertainty assessment or the simulation.

The hierarchical scheme described in Section 4.2 leads to multiple realizations of the study area under investigation. A variety of techniques, including indicator techniques, are used at different steps to arrive at a set of realizations that quantify the uncertainty. Each realization is a full specification of the study area: location, geometry, internal facies and petrophysical properties. These realizations must be post processed.

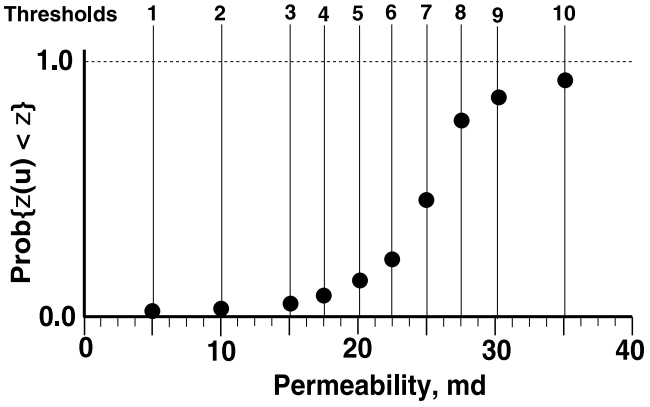


Figure 4. Example of a probability distribution derived from the indicator formalism. Each probability estimate is derived from kriging the data coded at that threshold.

4.6 Post Processing

Geostatistical models are useful for many purposes. The estimates at unsampled locations can be used directly for some decisions. The local uncertainty, that is, uncertainty at one location at a time is easily assembled from the multiple realizations. Maps can be made of P10 low values (the 0.1 quantiles of the local distributions) and the P90 high values (the 0.9 quantiles). These maps reveal two important features: (1) when the P10 value is high, then the actual value is surely high–there is a 90% probability to be even higher, and (2) when the P90 value is low, then the actual value is surely low–there is a 90% probability to be even lower. The local conditional variance could also be mapped to summarize local uncertainty. Another summary statistic that can be useful is the probability for the true value to be within a percentage (say 15%) of the estimate. There is little uncertainty when this probability is high.

Local estimates and local uncertainty are useful; however, they do not tell us the uncertainty at multiple locations simultaneously. Whenever uncertainty at many locations is required, then simulated realizations must be used.

Estimates are smooth and often inappropriate for direct input into flow simulation; flow predictions are biased because the connectivity of high permeability flow conduits and low permeability flow baffles is not accounted for. Simulated realizations are more appropriate. They also carry a measure of uncertainty. The paradigm of probabilistic analysis is that the set realizations are processed through a transfer function to assess uncertainty in response variables, see Figure 5.

Some response variables are straightforward such as volumetric calculation of resources. In fact, the response of smooth estimated models should match the average of the simulated realizations. In most cases, the transfer function is non-linear. Resources above a critical threshold and the response of flow and transport modeling are non-linear transfer functions. The response of an average is not the same as the average response.

All realizations are processed through the transfer function. This provides a distribution of uncertainty in the response variables. There are times when the transfer function (flow simulation) is very CPU-demanding. Moreover, many different scenarios of the transfer function must be considered. It is intractable to process all of the realizations through all scenarios.

The realizations are ranked according to some easy to calculate statistic such as the connected resource. Then, selected *low*, *median*, and *high* realizations are processed through the full transfer function. The ranking measure may be as simple as pore volume or as complex as the results of a fast flow

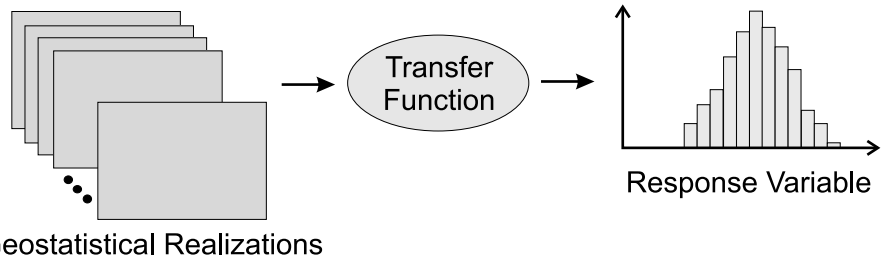


Figure 5. Schematic of how multiple realizations are processed through a transfer function to calculate uncertainty in a response variable (or multiple response variables).

simulation such as streamlines. A ranking measure of intermediate complexity often suffices: the connected volume to well locations is a good intermediate measure.

The preceding discussion has focused on uncertainty. Another aspect of post processing is sensitivity analysis, that is, determining how sensitive the response variables are to each of the input parameters/variables. This is done by holding some parameters constant or with experimental design techniques.

5. FUTURE DIRECTIONS

Geostatistical techniques for data fusion are applied in subsurface hydrology and other areas of geological modeling. A number of alternatives exist; however, the classical geostatistical paradigm presented here has had a history of successful prediction, is applied regularly and will provide unquestioned value in future applications.

The main problems with the geostatistical approach are that (1) it is poorly constrained by geological knowledge and processes, and (2) many statistical parameters must be inferred. No approach is perfect and people often want a change from the tried, true and boring applications with conventional techniques Alternatives are under consideration. Process-mimicking geological modeling, multiple point statistics, and data integration techniques provide interesting future directions.

Acknowledgments. Canada's NSERC organization and the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta are gratefully acknowledged for financial assistance.

REFERENCES

Chiles, J. P., and Delfiner, P. (1999). "Geostatistics : Modeling Spatial Uncertainty (Wiley Series in Probability and Statistics. Applied Probability and Statistics.)," Wiley, New York.

Cressie, N. (1991). "Statistics for Spatial Data." Wiley, New York.

David, M. (1977). "Geostatistical Ore Reserve Estimation." Elsevier, Amsterdam.

Deutsch, C. V., and Journel, A.G. (1997). "GSLIB: Geostatistical Software Library." Second Edition, Oxford University Press, New York.

Deutsch, C. V., (2003). "Geostatistical Reservoir Modeling," Oxford University Press, New York.

Goovaerts, P. (1997). "Geostatistics for Natural Resources Evaluation." Oxford University Press, New York.

Guardiano, F. and Srivastava, R. M. (1992), Multivariate Geostatistics: Beyond Bivariate Moments, in *Proceedings of Fourth International Geostatistics Congress*, Kluwer, New York.

Isaaks, E. H., and Srivastava, R. M. (1989). "An Introduction to Applied Geostatistics." Oxford University Press, New York.

Journel, A. G., and Huijbregts, C. (1978). "Mining Geostatistics." Academic Press, London.

Journel, A. G. (1983). Non-parametric estimation of spatial distributions, *Mathematical Geology*, 15(3), PAGES 445–469.

Kitanidis, P. (1997), *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press, New York.

Matheron, G. (1971). The theory of regionalized variables and its applications. Les cahiers du CMM. Fasc. No. 5, Ed. Ecole Nationale Superieure des Mines de Paris, Paris.

C. V. Deutsch, Department of Civil and Environmental Engineering, 3-40 NREF Building, University of Alberta, Edmonton, Alberta, CANADA T6G 2W2