Johan A. Huisman*
Jasper A. Vrugt
Ty P.A. Ferre

The guest editors introduce the papers that are featured in this special section and discuss how they contribute to addressing the current challenges in model–data fusion in the vadose zone.

J.A. Huisman, Agrosphere, IBG 3, Forschungszentrum Jülich, Jülich, Germany; J.A. Vrugt, Department of Civil and Environmental Engineering, University of California, Davis, CA, USA and Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, the Netherlands, and T.P.A. Ferre, Hydrology and Water Resources, University of Arizona, Tucson, AZ, USA. *Corresponding author (s.huisman@fz-juelich.de).

Vadose Zone Model–Data Fusion: State of the Art and Future Challenges

Models are quantitative formulations of assumptions regarding key physical processes, their mathematical representations, and site-specific relevant properties at a particular scale of analysis. Models are fused with data in a two-way process that uses information contained in observational data to refine models and the context provided by models to improve information extraction from observational data. This process of model–data fusion leads to improved understanding of hydrological processes by providing improved estimates of parameters, fluxes, and states of the vadose zone system of interest, as well as of the associated uncertainties of these values. Notwithstanding recent progress, there are still numerous challenges associated with model–data fusion, including: (i) dealing with the increasing complexity of models, (ii) considering new and typically indirect measurements, and (iii) quantifying uncertainty. This special section presents nine contributions that address the state of the art of model–data fusion.

Abbreviations: AFHO, actively heated fiber optics; GPR, ground penetrating radar; MCMC, Markov Chain Monte Carlo.

The past two decades have witnessed significant advances in vadose zone modeling and measurement technologies that have allowed the vadose zone community to tackle more complex problems with increasingly sophisticated measurement technologies. The required fusion of models with data is ideally achieved in a two-way process. Information contained in observational data is used to refine models, and the context provided by models is used to improve information extraction from available observational data or to identify information-rich data worth collecting. Despite the availability of more and better measurements and continued increases in computational power, it has become apparent that these advances have not solved the difficulties associated with such model–data fusion. Rather, these new capabilities have highlighted some of the more fundamental challenges common to all scientific analysis and challenged our approaches to model conceptualization, parameterization, validation, and hypothesis (model) reformulation. Specifically, they have tested our assumptions regarding the interpretation and value of observations, especially for indirect observations in complex environments. In addition, more complex models have encouraged vadose zone hydrologists to tackle problems with high parameter dimensionality and underdetermined inverse problems, which have shed light on shortcomings of our standard approaches for model parameterization that were not evident for less difficult problems. To address these shortcomings, there is a need for formal statistical methods that recognize the role of forcing data and model structural error in the analysis of parameter and predictive uncertainty.

There are numerous challenges in model–data fusion in vadose zone hydrology (e.g., see the review of Vrugt et al., 2008a). This special section focuses on three of the most pressing issues: model complexity, information extraction, and uncertainty quantification.

There is increasing use of complex highly parameterized models to describe coupled processes at scales large enough to support management and policy development. This use of increasingly complex models is associated with increases in parameter dimensionality and model run time, which put a premium on efficient methods for parameter estimation. In addition, the use of such coupled models requires multiple observation types, which implies the use of more advanced methods for model parameterization and information extraction from observations.

Most measurement methods available in vadose zone hydrology fall within one of three categories. Traditional, destructive measurements offer observations with high spatial

resolution but low temporal resolution. Automated monitoring stations typically provide measurements with high temporal resolution for a few selected points in space. Finally, geophysical and remotely sensed observations offer some mix of temporal and spatial resolution, but at the cost of being less direct measurements of hydrologically relevant properties (see the review by Robinson et al., 2008). There is growing recognition that it is not sufficient to collect data without careful consideration of the information necessary to discriminate among different model conceptualizations and parameterizations. Addressing this perennial problem will require continued advances in measurement technologies together with improved methods to identify those measurements that are most likely to provide the information that is most valuable for hypothesis testing and uncertainty reduction. This challenge of optimal data collection requires considerations of practicality and cost, as well as more specific considerations of how to reconcile typically conflicting information from different data types (e.g., Gupta et al., 1998), and how to consider data with varying spatial support (e.g., Hinnell et al., 2010; Huisman et al., 2010).

Model–data fusion offers opportunities for quantifying uncertainty in process understanding and hydrologic predictions. Methods that attempt to quantify uncertainty (Beven and Freer, 2001; Vrugt et al., 2003) are being used increasingly to treat the considerable uncertainty associated with vadose zone models. In particular, these methods have greatly improved our ability to identify and quantify uncertainty associated with model parameters. Moving forward, methods will have to consider uncertainty attributed to incomplete knowledge about model parameters, observational data of system inputs and outputs (Kuczera et al., 2006; Vrugt et al., 2008b), and model structure (e.g., Doherty and Welter, 2010; Gupta et al., 2012). Eventually, these assessments will have to encompass sources of uncertainty that are due to approaches and tools that are used for model–data fusion. In each case, it is likely that examinations of uncertainty will both improve our predictions and point to new directions for fundamental improvement in vadose zone hydrologic analysis.

Contents of the Special Section

We have provided a brief overview of the promises of model–data fusion and the challenges that this field faces for vadose zone applications. This special section presents nine contributions that illustrate ongoing efforts to address these challenges. We now briefly summarize each of these individual papers. For convenience, we have organized them along four main themes: (i) model–data fusion: the state of the art; (ii) model complexity; (iii) information extraction from new measurements technologies; and (iv) uncertainty quantification.

Model–Data Fusion: The State of the Art

A challenging issue in model–data fusion is whether the information content of the data is sufficient to obtain reliable estimates of model parameters. This issue was explored by Schelle et al. (2012), who investigated the amount of information that is required from a weighable lysimeter experiment to warrant the simultaneous identification of soil hydraulic and root distribution parameters of a vadose zone water flow model. Using in-silico (computer) experiments, they found that transient average water content and lysimeter outflow data are necessary to constrain the model parameters for a homogeneous soil. For a two-layer soil, additional matric potential measurements in both layers were required for a reliable model parameterization.

Keim et al. (2012) used a combination of measurements and modeling to derive a conceptual model for flow processes in unsaturated fractured rock. A numerical vadose zone flow model was first parameterized to predict winter drainage from the soil profile into the fractured rock using continuous measurements of soil water content and matric potential. These drainage estimates were compared with discharge measurements from fractures in a tunnel up to 45 m below the ground surface. The results clearly demonstrated that flow pathways converged with depth, which was additionally supported by the relatively short lag time between soil drainage and tunnel discharge.

Botros et al. (2012) developed six different numerical vadose zone models with varying levels of spatial detail and complexity of subsurface heterogeneity to predict vadose zone nitrate storage after a 7-yr-long fertilization experiment. Results demonstrated that all models consistently overestimated the actual nitrate storage. This finding points to incomplete knowledge of our understanding of the fate of nitrate in the vadose zone. Key controlling factors such as chemical heterogeneity related to mobile/immobile domains need to be explicitly considered to improve predictions of vadose zone nitrate flow and storage.

Summarizing, these contributions highlight some of the underlying challenges to model–data fusion. But, they also point to the promise of joint consideration of models, constrained on data, and data, interpreted in the context of models. In each case, we see that both models and data collection can be improved through model–data fusion.

Model Complexity

Verbist et al. (2012) reported on the successful parameterization of a coupled three-dimensional surface–subsurface model that describes runoff generation and vadose zone water flow. A global sensitivity analysis was used to determine the most sensitive model parameters, which were subsequently calibrated to observed soil water content and runoff data obtained during a high-intensity rainfall simulation experiment. An excellent match between the models and data was observed, inspiring confidence in the ability of a fully deterministic model to describe soil moisture dynamics and surface runoff.

Pagès et al. (2012) introduced a model to represent the architecture of a plant root system. Because of the stochastic nature of this model, it is difficult to estimate the model parameters directly from observations. Therefore, a large data set of root architectures was generated using forward simulation with model parameters drawn from prior parameter ranges. Global sensitivity analysis was subsequently used to create a statistical meta-model that directly related the model parameters to the root length density profiles of the simulated root architectures. Inversion of this meta-model showed that not all of the model parameters describing plant root architecture could be adequately estimated from the root length density profiles. Additional data beyond root length density were required to estimate the remaining model parameters.

These two contributions highlight the challenges of dealing with higher dimensionality problems, especially for highly nonlinear vadose zone processes. Model–data fusion is critical to addressing these challenges. Specifically, it is only through a joint consideration of models and data that we can determine the appropriate level of complexity that is supported by the existing data and that we can identify the data necessary to support a desired level of model complexity.

Information Extraction from New Measurement Technologies

It is widely recognized that vadose zone systems are highly heterogeneous. There are two approaches to deal with this heterogeneity. First, we can develop measurement technologies that provide measurements corresponding with the scales of heterogeneity. Second, we can develop measurement methods that "naturally average" properties of interest over larger scales. The difficulties inherent in collecting ubiquitous measurements are well known (e.g., Robinson et al., 2008; Vereecken et al., 2008). The challenges of interpreting large-scale measurements in highly heterogeneous conditions have not been explored in similar detail. Model–data fusion techniques are necessary to address the complex relationship between the measurement response and the heterogeneous distributions of parameters, states, and fluxes.

Jadoon et al. (2012) presented an integrated or coupled hydrogeophysical inversion approach to directly estimate vadose zone soil hydraulic properties from observations of time-lapse ground penetrating radar (GPR). The complex relationship between the modeled soil moisture states and the GPR measurements was accounted for by coupling a forward model of the measurement process to a vadose zone model. The use of a dual-porosity model parameterization improved the reliability of the estimated soil hydraulic parameters and resulted in a better representation of near-surface moisture content profiles.

Actively heated fiber optics (AFHO) is an emerging method for determining distributed soil moisture over distances up to several tens to hundreds of meters. In this method, the metal sheath surrounding the fiber optic cable is used to generate a heat pulse, and the cooling of the soil after the heat pulse was monitored using a distributed temperature sensing system. An analytical solution describing radial heat flow was used to estimate the soil thermal conductivity, which can be converted to moisture content using a soil-specific calibration. Ciocca et al. (2012) evaluated this new approach in a lysimeter study and obtained an adequate agreement with reference moisture content measurements. Their analysis also indicated that the use of longer heat pulses and increased temporal sampling further improved the accuracy of the AFHO-observed soil water contents.

Considered together, these contributions point out that, while it can be difficult to interpret large scale measurements in the vadose zone, model–data fusion provides the best approach to quantitatively combining understanding of instrument responses, spatial heterogeneity, and hydrologic processes.

Uncertainty Quantification

Analyses of vadose zone hydrologic processes entail considerable uncertainty, and it is of eminent importance to account for this uncertainty to improve process understanding and model predictions. Model–data fusion does not solve this problem outright, but it does provide an objective and consistent tool for examining all sources of this uncertainty in a common framework. In this special section, the contribution of Shi et al. (2012) compares nonlinear regression and Markov Chain Monte Carlo (MCMC) simulation to evaluate the predictive performance of a two-phase flow model using cross-validation. It was found that MCMC methods provide a more accurate representation of predictive uncertainty. In addition, their modeling results dispute the common notion that nonlinear regression techniques are computationally superior to state-of-the-art MCMC sampling algorithms. The power of MCMC analysis is further illustrated by Scholer et al. (2012). They explored the information content of time-lapse ground penetrating radar measurements made during a dynamic vadose zone infiltration experiment to constrain the estimation of hydraulic parameters. The MCMC analysis showed that the posterior uncertainty in the model parameters was considerably reduced compared to their prior ranges. Looking forward, it will be important to bring these analyses "full circle" and to use model–data fusion to identify measurements that are best able to reduce existing uncertainties in the context of the hydrologic models.

Summary and Outlook

Progress in model–data fusion can be assessed by comparing the scope and content of this special section with that of its predecessor in *Vadose Zone Journal* (Vrugt and Neuman, 2006). Over the intervening 6 years, the evaluation of model parameter uncertainty has become more established (Schelle et al., 2012; Shi et al., 2012; Scholer et al., 2012), and more advanced measurement technologies are being used for model conceptualization and

parameterization (Ciocca et al., 2012, Keim et al., 2012; Jadoon et al., 2012). Three studies explored the effect of model structural adequacy on model–data fusion results (Botros et al., 2012; Jadoon et al., 2012; Schelle et al., 2012).

Despite this progress, we do feel that the vadose zone community has not yet fully recognized that the results of model–data fusion to a large extent depend on the appropriate and simultaneous treatment of uncertainty in input data, output data, model parameters, and model structure. This may be partly explained by the high computational effort required for vadose zone models, which makes them less amenable to computationally intensive treatments of uncertainty. Nevertheless, we feel that the time has come to turn our community's attention to addressing the underlying challenges of uncertainty quantification and reduction. We strongly encourage a dialog with other communities facing this same challenge along the lines proposed by Gupta et al. (2012). Finally, we remain confident that approaches based on model–data fusion provide the most promising path forward to resolve model structural inadequacies, improve model calibration, and identify information-rich data for collection.

Acknowledgments

References
Beven, K., and J. Freer. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology. J. Hydrol. 249:11–29. doi:10.1016/S0022-1694(01)00421-8

Botros, F.E., Y.S. Onsoy, T.R. Ginn and T. Harter. 2012. Richards equation–based modeling to estimate flow and nitrate transport in a deep alluvial vadose zone. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0145

Ciocca, F., I. Lunati, N. van de Giesen, and M.B. Parlange. 2012. Heated optical fiber for distributed soil-moisture measurements: A lysimeter experiment. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0177.

Doherty, J., and D. Welter. 2010. A short exploration of structural noise. Water Resour. Res. 46:W05525. doi:10.1029/2009WR008377

Gupta, H.V., S. Sorooshian, and P.O. Yapo. 1998. Towards improved calibration of hydrologic models: Multiple and non-commensurable measures of information. Water Resour. Res. 34:751–763. doi:10.1029/97WR03495

Gupta, H.V., M.P. Clark, J.A. Vrugt, G. Abramowitz, and M. Ye. 2012. Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 48:W08301. doi:10.1029/2011WR011044

Hinnell, A.C., T.P.A. Ferré, J.A. Vrugt, J.A. Huisman, S. Moysey, J. Rings, and M.B. Kowalsky. 2010. Improved extraction of hydrologic information from geophysical data through coupled hydrogeophysical inversion. Water Resour. Res. 46:W00D40. doi:10.1029/2008WR007060

Huisman, J.A., J. Rings, J.A. Vrugt, J. Sorg and H. Vereecken. 2010. Hydraulic properties of a model dike from coupled Bayesian and multi-criteria hydrogeophysical inversion. J. Hydrol. 380:62–73. doi:10.1016/j.jhydrol.2009.10.023

Jadoon, K.Z., L. Weihermüller, B. Scharnagl, M.B. Kowalsky, M. Bechtold, S.S. Hubbard, H. Vereecken, and S. Lambot. 2012. Estimation of soil hydraulic parameters in the field by integrated hydrogeophysical inversion of time-lapse ground-penetrating radar data. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0177.

Keim, D.M., L.J. West, N.E. Odling. 2012. Convergent flow in unsaturated fractured chalk. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0146

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer. 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. J. Hydrol. 331:161–177. doi:10.1016/j.jhydrol.2006.05.010

Pagès, L., C. Bruchou, and S. Garré. 2012. Links between root length density profiles and models of the root system architecture. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0152.

Robinson, D.A., C.S. Campbell, J.W. Hopmans, B.K. Hornbuckle, S.B. Jones, R. Knight, F. Ogden, J. Selker, and O. Wendroth. 2008. Soil moisture measurement for ecological and hydrological watershed scale observatories: A review. Vadose Zone J. 7:358–389. doi:10.2136/vzj2007.0143

Schelle, H., S.C. Iden, J. Fank, and W. Durner. 2012. Inverse estimation of soil hydraulic and root distribution parameters from lysimeter data. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0169.

Scholer, M., J. Irving, M. Looms, L. Nielsen, and K. Holliger. 2012. Bayesian Markov-Chain-Monte-Carlo inversion of time-lapse crosshole GPR data to characterize the vadose zone at the Arrenaes site, Denmark. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0153.

Shi, X., M. Ye, S. Finsterle, and J. Wu. 2012. Comparing nonlinear regression and Markov Chain Monte Carlo methods for assessment of prediction uncertainty in vadose zone modeling. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0147.

Verbist, K., S. Pierreux, W.M. Cornelis, R. McLaren, and D. Gabriels. 2012. Parameterizing a coupled surface–subsurface three-dimensional soil hydrological model to evaluate the efficiency of a runoff water harvesting technique. Vadose Zone J. 11 (this issue). doi:10.2136/vzj2011.0141.

Vereecken, H., J.A. Huisman, H. Bogena, J. Vanderborght, J.A. Vrugt and J.W. Hopmans. 2008. On the value of soil moisture measurements in vadose zone hydrology: A review. Water Resour. Res. 44:W00D06. doi:10.1029/2008WR006829

Vrugt, J.A., H.V. Gupta, W. Bouten, S. Sorooshian. 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour. Res. 39:1201. doi:10.1029/2002WR001642

Vrugt, J.A. and S.P. Neuman. 2006. Introduction to the special section in Vadose Zone Journal: Parameter identification and uncertainty assessment in the unsaturated zone. Vadose Zone J. 5:915–916. doi:10.2136/vzj2006.0098

Vrugt, J.A., P.H. Stauffer, T. Wöhling, B.A. Robinson, and V.V. Vesselinov. 2008a. Inverse modeling of subsurface flow and transport properties: A review with new developments. Vadose Zone J. 7:843–864. doi:10.2136/vzj2007.0078

Vrugt, J.A., C.J.F. ter Braak, M.P. Clark, J.M. Hyman, and B.A. Robinson. 2008b. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour. Res. 44:W00B09. doi:10.1029/2007WR006720