Multi-Model Validation Assessment of Groundwater Flow Simulation Models Using Area Metric Approach

Omkar Aphale and David J. Tonjes,
Department of Technology and Society, Stony Brook University

Introduction

Ignoring model uncertainties, epistemic and/or aleatory, may overestimate the goodness-of-fit or, validity of a groundwater model. As a remedy, multiple models should be developed and their validity be assessed. Typically, validation involves quantifying the level of agreement between the deterministic observed data and their corresponding model-simulated values. Here, an alternative, non-deterministic approach called the "area metric" was adopted for the validation assessment. This metric quantifies the level of disagreement between the distributions of the observed data and that of the model-simulated values. The application of the area metric-based multi-model validation assessment was demonstrated using a case study of groundwater flow simulation model of the municipal landfill in the Town of Brookhaven, NY.

Model Uncertainties

Models offer a powerful, low-cost learning environment to test theories that enhance our understanding of the real-world systems (Bredehoeft, 2005). In the field of hydrogeology, models are commonly used either (i) to characterize the hydrogeologic regime of the study area, (ii) to augment conceptual understanding of its functioning, (iii) to organize the different types of data collected about the hydrogeologic system, (iv) to make predictions about its future behavior, and/ or (v) to gauge its response to changes in the normal conditions, to extreme or sudden stresses, or to remedial measures (Anderson and Woessner 1994, p. 4-5).

A key challenge in the modeling of groundwater flow systems is to deal with the uncertainties associated with their configuration. Model uncertainty is not only associated with the inputs and the parameter values that enter into a model, but also with the model's conceptual and mathematical structure (Neuman, 2003). Likewise, model uncertainty is divided into three classes – input uncertainty, parameter uncertainty, and conceptual uncertainty (Beven, 2012; Refsgaard et al., 2006). Another type of classification is based on the nature of uncertainty and it focuses on the thematic causes of the rise of different uncertainties. Here, the model uncertainty is divided into two classes – epistemic uncertainty and aleatory uncertainty.

1. Epistemic uncertainty (EU) arises because of the absence or incompleteness of knowledge about the characteristics and the behavior of hydrogeologic system being modeled. The knowledge deficiency could result from measurement uncertainty, non-detects, data censoring, missing values, use of surrogate data, imperfections in scientific understanding, rounding error, intermittent measurement of periodic processes, subjective judgments, and / or, ambiguities (Oberkampf et al., 2002).

2. Aleatory uncertainty (AU) arises because of the effect of chance and it is a function of natural stochasticity of the system. The natural stochasticity in the system could result from the inherent variability of the system, or environmental or structural variations across space

or thorough time, or heterogeneity among components, external input data and functions, parameters, and / or model structures (Oberkampf et al., 2002).

Multi-Model Validation Assessment

Given model uncertainties, restricting the modeling exercise to a fixed, singular model limits the model's ability as a decision-support tool. Instead, multiple models, based on varying combinations of the inputs, parameters, and conceptualizations should be developed and evaluated. This reduces the chances of model over-fitting that, in turn, reduces the chances of rejecting a valid model (Type I error) as well as the chances of failing to reject an invalid model (Type II error) (Neuman and Wierenga, 2003).

Given a set of multiple models, a modeler can perform "multi-model analysis" to evaluate the models' goodness-of-fit, to the real-world system or, their representativeness. The process of evaluation of a model's representativeness is generally referred to as model "validation" (Law and Kelton, 2000, p. 264). The simplest form of model validation is to measure the level of agreement between the scalar, deterministic, observed data and their corresponding model-simulated values. Different multi-model analyses, such as information criteria-based techniques (Poeter and Anderson, 2005), multi-model averaging (Ye et al., 2010), multi-objective optimization (Yapo et al., 1998), multi-objective clustering (Handl and Knowles, 2005), and Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Freer, 2001) have been adopted in the literature.

An alternative validation assessment approach called the "area metric" facilitates non-deterministic validation assessment. In this approach, the level of disagreement is calculated between the "distributions", specifically, the empirical cumulative distribution function (ECDF) derived from the observed data and the ECDF derived from the model-simulated values. For instance, Figure 1 shows four ECDFs; three represent the ECDFs derived from the model-simulated data from three different models of the same system, while one ECDF (with solid circles) represent the observed data. Notice that the model ECDFs occupy different positions on the horizontal axis; some are placed closer to the observed data ECDF than others. Correspondingly, the value of the area metric (A), that is, the area of the space between a model ECDF and the observed data ECDF is different for different models. In this figure, $A_{\text{Model 1}} < A_{\text{Model 2}} < A_{\text{Model 3}}$.

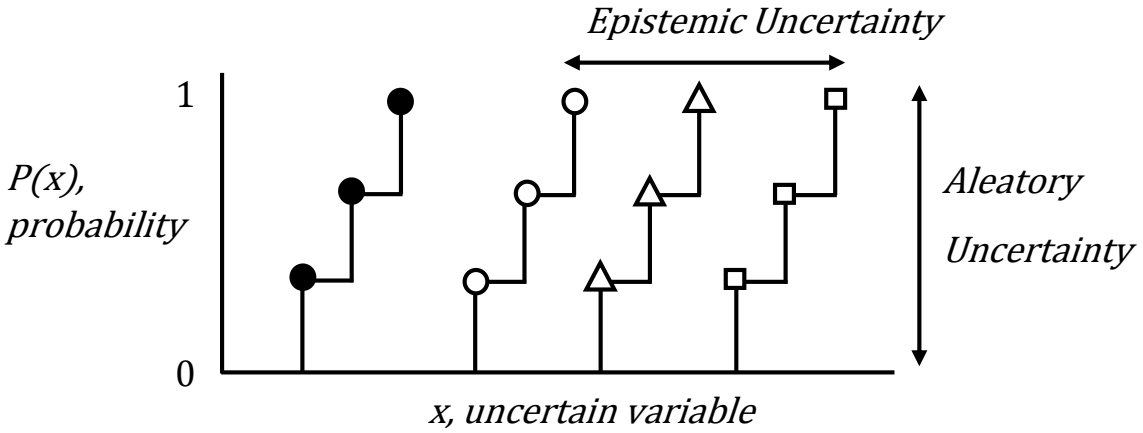

Figure 1: The ECDF of the observed data (solid circle) and the ECDF of the simulated output from 3 distinct models; Model 1 is represented by hollow circles, Model 2 by hollow triangles, and Model 3 by hollow squares

The area metric approach disaggregates both types of model uncertainties, epistemic and aleatory. The spread of distributions along the horizontal axis represents epistemic uncertainty, while the dispersion within a given ECDF represents aleatory uncertainty; the aleatory uncertainty is embedded within each distinct ECDF, as seen in the figure above. The area metric is mathematically well-behaved and is expressed in the same units as that of the observed data (Ferson et al., 2008).

Objective and Scope

I adopted an area matric-based multi-model validation assessment to evaluate the representativeness of groundwater flow simulation models. I hypothesized that this approach facilitates a robust multi-model analysis that allows selection of, from a model space, those models that are better representations of the real-world groundwater flow system. The application of this approach was demonstrated for a case study of groundwater flow simulation model being developed for the municipal landfill site in the Town of Brookhaven in Suffolk County, New York ("the landfill model").

Methods

361 different versions of the landfill model, with varying model configurations, were generated. Each model was a unique combination nine, pre-selected variable model features (Table 1). 8 of these variable features represented epistemic model uncertainties. The variations in these features were represented by different states, usually two or three in number. The remaining variable feature represented the aleatory uncertainty associated with the landfill model, that is, it represented the natural fluctuation of water table from high, to median, to low levels. Explanation of these variables was beyond the scope of this study.

Code	Variable Feature	State 1	State 2	State 3
A	Bottom of layer 1	Uniform thickness	Interpolated surface	--
B	Bottom of layer 2	Constant slope	Interpolated surface	--
C	Extent of the PSU	2-zone	3-zone	--
D	Recharge (regional)	Yes	No	--
E	Recharge (local)	Natural	Via Recharge Basins	Zero recharge
F	Stream segmentation	Yes	No	--
G	Kh – UGA (ft/d)	High	Medium	Low
	L1	300	250	200
	L2	250	200	150
	L3	200	150	100
H*	Constant head boundary at the northern edge of the model (feet)	42	40	38
I	Top surface of the PSU	Constant slope	Interpolated surface	--

Table 1: Variable features and their states (*represents aleatory uncertainty; UGA = Upper Glacial aquifer, PSU = potentially semi-confining unit)

The models were simulated as three dimensional, steady-state, finite-difference groundwater flow simulations using Visual MODFLOW v. 4.2.

The observed data was derived from the 133 head observation wells distributed across the study area and screened in different aquifer units (the Upper Glacial aquifer or the Magothy aquifer) at different depths. The ECDFs of the observed data were discrete step functions with three steps. These steps represented the maximum, the median, and the minimum head observations made at each well. 133 observation data ECDFs were generated in this manner.

The simulated data ECDF was generated by simulating each model three times for three water table conditions – high, median, and low. The model configurations remained fixed, only the values of the aleatory variable (H) were changed from 42’ to 40’ to 38’ feet to represent the fluctuation in the water table. Three simulated head values were generated upon simulation, one per model iteration. These values were then collated to form the simulated data ECDF for a given well. This process was repeated for all 133 head observation wells. 133 simulated data ECDFs, one for each observation well, were generated in this manner.

The value of the area metric (A) was calculated by comparing the observed data ECDF with its corresponding simulated data ECDF for a given well. This process was repeated for each of the 133 individual head observation wells for a given model. The resultant 133 A values were then collated into a model ECDF. Each model ECDF was compared with a reference model ECDF to generate an overall area metric (A*) value (the reference model is a hypothetical model where A= 0 for all the head observation wells). This process was repeated for all 361 models to generate 361 A* values. The models were then arranged in an ascending order (from smallest to largest) of their A* values.

Results

The results were graphically summarized in a meta-distribution whose elements are the 362 model ECDFs, including 361 model-ECDFs (in black) and the reference ECDF depicted as a spike distribution at 0 feet (in red) (Figure 2a). Each model-ECDF is generated from 133 A values calculated of the 133 head observation wells for the given model. 159 models were rejected due to the violation of condition of monotonicity, while 202 models were retained for further analysis. Discussion on montonicity is beyond the scope of this paper. The following results are derived from the retained (202) models. For example, Figure 2b shows the revised meta-distribution after the removal of rejected models' ECDFs. It can be observed that certain model ECDFs appear to be closer to the reference ECDF. Also, the dispersion within an individual model ECDF indicates that different A values were calculated for different head observation wells.

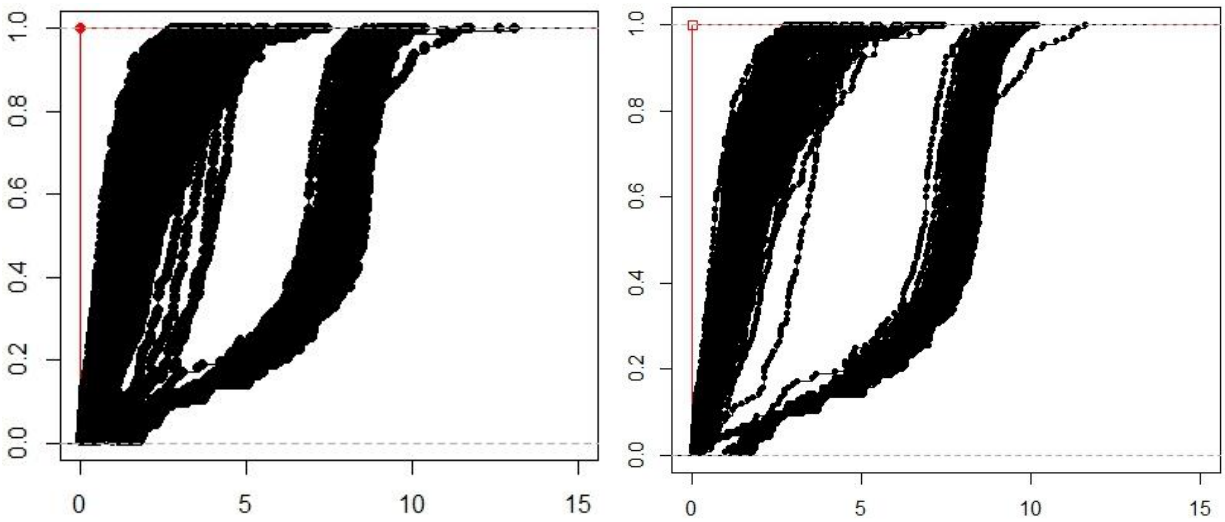

Figure 2: Meta-distribution of (a) the model ECDFs (in black) and the reference model (in red), (b) revised meta-distribution (horizontal axis in feet, vertical axis is cumulative probability)

The smallest A* value was 0.79 feet (model #266), while the largest A *value was 7.39 feet (model #287). The median A* value was 1.66 feet, the mean A* value was 3.17 feet with one standard deviation of 2.55 feet. Models with smaller A* value show less disagreement with the observed data and therefore were deemed to have higher validity than the models with larger A* values.

Figure 3 shows the vertical distribution of the average of the A values with respect to the screen depth of the head observation wells. The figure shows that the average A* values of the shallow wells were spread between 0 and 5 feet with clustering of values between 3 and 4 feet. The intermediate and deep wells had average A values between 2 and 4 feet. 70 of the 133 HOB wells were screened at the depths between 20 and -20 feet to the mean sea level (msl), and a dense cluster of values was observed at those depths. This figure indicated that the A values showed a noticeable geo-spatial distribution along the vertical extent of the model domain.

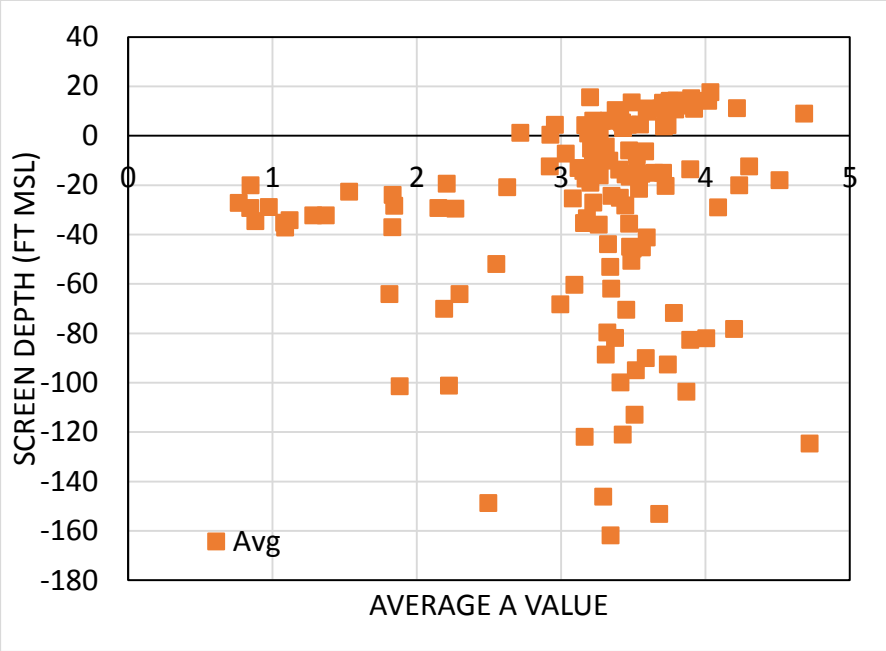

Figure 3: Vertical distribution of the average A values relative to well depths (values in feet)

The differences between the A* values of different states of a given variable were statistically compared using one factor unbalanced ANOVA. For example, out of 202 models, 133 models contained state 1 of variable feature A (A1) in their configuration, while 69 models contained state 2 of the same variable (A2). The one-way ANOVA indicated that the difference between the A* values of the A1 models and the A2 models was not statistically significant. The test indicated that the difference between/among the A* values were statistically significant for the variable features B, D, and E (Table 2). This suggests that certain state-variable feature combinations influenced the representativeness of the model more than other combinations.

Sates of variable features	F value	Pr (>F)
A1A2	0.793	0.3743
B1B2	9.3931	0.002479***
C1C2	1.3803	0.2415
D1D2	2638.2	2.2e-16***
E1E2E3	28.291	1.53e-11***
F1F2	1.1236	0.2904
G1G2G3	0.9196	0.4004
I1I2	0.7059	0.4018

Table 2: Results of the one-way unbalanced ANOVA (*= p<0.5, **=p<0.01, ***=p<0.0001)

Table 3 shows the model configurations of the top 10 models whose A* values were the smallest. Table 4 shows the sort of these top 10 models according to their constituent state-variable feature combinations for each variable feature. For example, 3 out of the top 10 models contained state 1 of the variable feature A, while 7 models contained state 2 of the variable feature. It was found that – A2 > A1 models, C2 > C1 models, D1 >> D2 models, and F1 > F2

models. No noticeable difference was found in other state-variables. This result suggests that certain state-variable feature combinations appeared more commonly in the top 10 models than certain other combinations. This is useful in streamlining the pool of potential model configurations.

Model	Variable feature combinations								A* (ft)
266	A1	B1	C1	D1	E1	F1	G1	I1	0.7907
252	A2	B2	C2	D1	E3	F1	G3	I1	0.9226
240	A2	B2	C2	D1	E2	F1	G3	I1	0.9279
257	A2	B2	C2	D1	E3	F2	G2	I2	0.9493
245	A2	B2	C2	D1	E2	F2	G2	I2	0.9494
228	A2	B2	C2	D1	E1	F1	G3	I1	0.9570
154	A2	B1	C2	D1	E1	F1	G1	I2	0.9747
49	A1	B1	C2	D1	E2	F1	G1	I1	0.9966
27	A1	B1	C1	D1	E3	F1	G2	I1	0.9972
172	A2	B1	C2	D1	E2	F2	G1	I2	0.9992

Table 3: Configuration of the top 10 models with smallest A* values

Variable feature	States of the variable feature		
	State 1	State 2	State 3
A	3	7	--
B	5	5	--
C	2	8	--
D	10	0	--
E	3	4	3
F	7	3	--
G	4	3	4
I	6	4	--

Table 4: Sorting of the top 10 models on the basis of their constituent state-variable feature combinations for each variable feature

Discussion

A common drawback of the traditional model validation methods is that they assume that the simulated and the observed data are deterministic quantities that are devoid of uncertainties. In the above approach, instead of making this unjustifiable assumption, model uncertainty has been explicitly incorporated using multiple model conceptualizations and the representativeness of these multiple models is tested over a range of observed data in a tangible manner. In addition, the area metric approach acts as a confidence building exercise that helps to reduce the model users' uncertainty about the usefulness of the model because smaller values of the area metric increase our confidence about the operational-replicative validity of the model that, in turn, increases our confidence about the conceptual validity of the model. Also, the area metric approach enables the modeler to calibrate the concept rather than the commonly used, one-

dimensional model calibration approach that can only calibrate the parameter to achieve a close agreement between the observed and the simulated data.

The area matric does not indicate absolute validity of the model and it is a measure of representativeness with respect to the application domain for which the model has been developed. Incomplete, infrequent, missing, and potentially erroneous data make it difficult to identify the unique model from the model space, even if theoretically such a model exists. Also, it is necessary to separate model's accuracy from the model's adequacy to maintain the objectivity of the area metric. The area metric approach is applied to a select pool of multiple models specified by the modeler. Therefore, the modelers' knowledge of the numerical methods and the hydrogeology of the study area are of seminal importance in configuring these models.

Future work includes (i) using probability bounds analysis (PBA) for further assessment of the results, (b) acknowledge and incorporate recognizable errors as models uncertainties, and (iii) developing multi-system response quantity validation assessment.

Acknowledgements

This project is being funded by the Division of Waste Management, Town of Brookhaven. We thank Ed Hubbard, Commissioner of Waste Management, Town of Brookhaven, for his encouragement for this project.

References

1. Anderson, M., and W. Woessner, 1992, Applied Groundwater Modeling: Simulation of Flow and Advective Transport, Volume 4, Academic Press, p. 4-5.
2. Beven, K., 2012, Causal Models as Multiple Working Hypotheses about Environmental Processes, Comptes Rendus Geoscience, 344(2), 77-88.
3. Beven, K., J. Freer, 2001, Equifinality, Data Assimilation, and Uncertainty Estimation in Mechanistic Modeling of Complex Environmental Systems Using the GLUE Methodology, Journal of Hydrology, 249(1), 11-29.
4. Bredehoeft, J., 2005, The Conceptualization Model Problem-Surprise, Journal of Hydrogeology, v. 13(1), p. 37-46.
5. Ferson, S., W.L. Oberkampf, and L. Ginzberg, 2008, Model Validation and Predictive Capability for the Thermal Challenge Problem, Computer Methods in Applied Mechanics and Engineering, v. 197(29), p. 2408-2430.
6. Handl, J., Knowles, J., 2005, Exploiting the trade-off: The benefits of MO in data clustering, EMO 2005, Coello, C.A., Coello et al. (Eds), LNCS 3410, pp. 547-560, Springer-Verlag, Heidelberg, Germany.
7. Konikow, L. F., 1996, Numerical Models of Groundwater Flow and Transport, in Manual on Mathematical Models in Isotope Hydrology, IAEA-TECDOC-910, International Atomic Energy Agency, Vienna, Austria.
8. Law, A., Kelton, W.D., 2000, Validation of Simulation Models, Simulation Modeling and Analysis, McGraw Hill, 3rd Ed., Boston, MA, p. 264.
9. Neuman, S., 2003, Maximum Likelihood Bayesian Averaging of Uncertain Model Predictions, Stochastic Environmental Research and Risk Assessment, v. 17(5), p. 291-305.

10. Neuman, S., P. Wierenga, 2003, A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites, NUREG/CR-6805, Washington, DC: U.S. Nuclear Regulatory Commission.
11. Oberkampf, W. L., S.M. DeLand, B.M. Rutherford, K.V. Diegert, and K.F. Alvin, 2002, Error and Uncertainty in Modeling and Simulation, Reliability Engineering and System Safety, v. 75(3), p. 333-357.
12. Poeter, E., D. Anderson, 2005, Multimodel Ranking and Ground Water Modeling, Ground Water, 43(4), 597-605.
13. Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., Troldborg, L., 2012, Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. Advances in Water Resources, 36, 36-50.
14. Refsgaard, J. C., J.P. Van der Sluijs, J. Brown, and P. Van der Keur, 2006, A Framework for Dealing with Uncertainty Due to Model Structure Error, Advances in Water Resources, v. 29(11), p. 1586-1597.
15. Romanowicz, R., and R., MacDonald, 2005, Modeling Uncertainty and Variability in Environmental Systems, Acta Geophysica Polonica, v. 53(4), p. 401-417.
16. Singh, A., S. Mishra, and G. Ruskauff, 2010, Model Averaging Techniques for Quantifying Conceptual Model Uncertainty, Groundwater, v. 48(5), p. 701-715.
17. Yapo, P., H. Gupta, S. Sorooshian, 1998, Multi-objective global optimization for hydrologic models, Journal of Hydrology, 204, 83-97.
18. Ye, M., K.F. Pohlmann, J.B. Chapman, G.M. Pohll, D.M. Reeves, 2010, A Model-Averaging Method for Assessing Groundwater Conceptual Model Uncertainty, Ground Water, 48(5), 716-728.