2017

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

Omkar Aphale
SUNY Stony Brook, omkar.aphale@gmail.com

David J. Tonjes
SUNY Stony Brook, david.tonjes@stonybrook.edu

Research Paper/

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

Omkar Aphale
Corresponding Author: 347A Harriman Hall, Stony Brook University, Stony Brook, NY 11794-3760; omkar.aphale@alumni.stonybrook.edu

David J. Tonjes
344 Harriman Hall, Stony Brook University, Stony Brook, NY 11794-3760;
david.tonjes@stonybrook.edu

Published as:
Aphale, O., and DJ Tonjes. 2017. Multi-model validation assessment of groundwater flow simulation models using area metric approach. Groundwater 55(2):219-226.*DOI 10.1111/gwat.12470*

Conflict of Interest: None

Key words: Area Metric, validation, multi-model, groundwater, epistemic, aleatory, uncertainty

Article Impact Statement: Area Metric approach incorporates epistemic and aleatory uncertainties and assesses multiple models' validity over a range of observed data.

Abstract

We demonstrate the application of the Area Metric developed by Ferson et al. (2008) for multi-model validity assessment. The Area Metric quantified the degree of models' replicative validity: the degree of agreement between the observed data and the corresponding simulated outputs represented as their empirical cumulative distribution functions (ECDFs). This approach was used to rank multiple representations of a case study groundwater flow model of a landfill by their Area Metric scores.

A multi-model approach allows for the accounting for uncertainties that may either be epistemic (from lack of knowledge), or aleatory (from variability inherent in the system). The Area Metric approach enabled explicit incorporation of model uncertainties, epistemic as well as

aleatory, into validation assessment. The proposed approach informs understanding of the collected data and that of the model domain. It avoids model overfitting to a particular system state, and in fact is a blind assessment of the models' validity: models are not adjusted, or updated, to improve their fit. This approach assesses the degree of models' validity, in place of the typical binary model validation/invalidation process. Collectively, this increases confidence in the model's representativeness that in turn, reduces risks to model users.

Introduction

Simply put, model validation is the process of assessing model representativeness. The exact definitions and the feasibility of model validation is widely debated (Bredehoeft 2003; Oreskes 1998; Oreskes et al. 1994). Here, we define validation as "replicative validation": quantifying agreement between observed data and corresponding simulated values.

Two practical constraints limit the deterministic (in)validation of a model. First, the observed data, such as groundwater heads vary over space and time and are not static. Further, a singular observed data set represents a "snapshot" of reality: a state of the system at an instance in time and space, instead of representing the range of the system's behavior. Practical tendency is to tune models to this snapshot, but such overfitted models tend to perform poorly when tested against data set from a different state (Konikow 1996). Also, an individual datum is seldom deterministic due to the associated measurement error of uncertain magnitude arising from either manual, technical, or recording errors (Romanowicz and MacDonald 2005).

Secondly, the simulated outputs are generated by a model, one that is the product of assumptions, simplifications, and lumped approximations (Waganer and Gupta 2005). Historical data and surrogate data are used as inputs, continuous terrains and geology are discretized into a finite-grid model domain, while time is aggregated into coarse steps (Refsgaard et al. 2012). Commonly heterogeneous aquifer property, such as the hydraulic conductivity, is lumped into a

single parameter value (Beven and Binley 1992). Consequently, although the model-simulated values are deterministic and have a one-on-one correspondence with observed data, the model may not replicate the exact state of the system when the observations were made (Beven 2012).

Thus, a key challenge in modeling is to deal with the uncertainty about the configuration of the system to be modeled. This uncertainty could be "epistemic", arising due to absence or incompleteness of our knowledge about the system, due to measurement error, non-detections, data censoring, missing values, use of surrogate data, or rounding error. Or, this uncertainty could be "aleatory", arising because of the natural stochasticity of the system, environmental or structural variations across space or thorough time, heterogeneity among components of the groundwater system and from external input data and functions, and parameterization (Oberkampf and Barone 2006). Given the uncertainty, correspondence between the model and the reality is unlikely to be exact.

As a remedy, multiple model depictions of varying inputs, parameters, and conceptualizations should be constructed. Subsequently, their validity be assessed to find those models that fit the reality better, instead of trying to achieve an exact fit to a singular model to a snapshot representation of reality. Approaches adopted in the past for multi-model analysis include information criteria-based model selection (Poeter and Anderson 2005), MMA (multi-model averaging; Singh et al. 2010), MOO (multi-objective optimization; Yapo et al. 1998), and GLUE (Generalized Likelihood Uncertainty Estimation; Beven and Binley 1992).

The objective of this paper is to demonstrate a multi-model validity assessment approach based the Area Metric, a performance indicator called, developed by Ferson et al. (2008). The concept of Area Metric-based validity assessment is primarily applied in risk assessment (Oberkampf and Barone 2006; Ferson and Tucker 2003). To our knowledge, the proposed

approach is a unique contribution to the extant array of techniques used to assess the validity of groundwater flow simulation models.

As case study, a simulation of groundwater flow near a landfill in New York, USA was used with the Area Metric applied to assess the replicative validity of multiple variants of a base model. The multi-model case study demonstrates that the proposed approach facilitates a robust multi-model analysis that identities those model variants that are better representations of the groundwater flow system.

Research Method

Area Metric

The Area Metric is defined here as the integral of the absolute value of the difference between the empirical cumulative distribution functions (ECDFs) generated from the observed data ($ECDF_{observed}$) and the ECDF generated from the model-simulated outputs ($ECDF_{simulated}$). An ECDF represents the cumulative probability that a variable X, such as the groundwater heads, will be less than or equal to different observed values, x_i (i=1,…,n), possible of X (Morgan and Henrion 2006, p. 74). An ECDF is a monotonically increasing discrete distribution ranging from probabilities 0 to 1 with n vertical steps of equal length. The ECDF is truncated with the finite interval ranges of x values. It may estimate the true CDF of X with a very large number of observations (Ferson et al. 2008).

The Area Metric is independent of the quantum of the observed data. The Area Metric is expressed in the same units as the observed data because the ECDF is plotted on a dimensionless (L^0) probability scale on the vertical axis while the observed data (L^1) plotted along the horizontal axis in increasing order of magnitude. The Area Metric can become mathematically analogous to other performance indicators, such as simple Euclidean distance, or the Mean Squared Error (MSE) depending on the constitutions of the ECDFs of the observed and/or the

simulated data. The Area Metric is non-parametric because no assumptions are necessary regarding the statistical nature of the observed and the simulated data (Roy and Oberkampf 2011; Ferson et al. 2008). Smaller Area Metric values describe better overall agreement between observed and simulated data, or, better replicative validity.

Case study models

The case study model simulates the groundwater flow in the vicinity of the municipal landfill site located in southeastern Suffolk County, New York. The model domain covered about 32 mi^2 (83 km^2) encompassing the landfill that is located about 12,000 feet (3.6 km) south of the regional ground water divide (Figure 1). The topography is flat, southward sloping, ranging from about 80 feet to the northwest to near sea level to the southeast. The principle axis of groundwater flow is southeasterly in the water table and underlying aquifers.

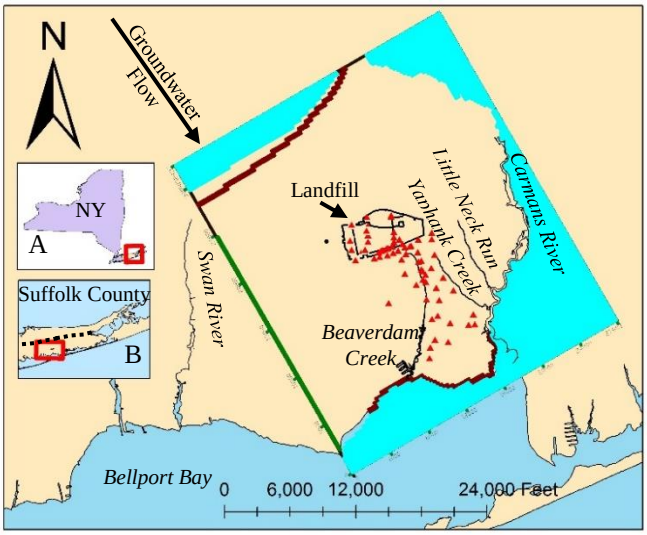

Figure 1: Landfill and vicinity, along with the New York State (inset A), and Suffolk County (inset B) with the regional ground water divide (dotted line). Map also shows the model domain with inactive zones (blue), GHB boundary (green), CHD boundary (brown), observation wells (red triangles), and the public supply well (black dot).

The model domain was bounded by constant head (CHD) boundaries representing the regional hydrologic divide to the northeast and the Bellport Bay to the south. The Swan River was simulated as a general head boundary (GHB) to the southwest. The Beaverdam Creek, Carmans River, Yaphank Creek, and Little Neck Run were simulated as drains because 95% of their baseflow is estimated to be groundwater (Peterson 1987).

The model domain was vertically discretized into five layers. The upper three layers (L1, L2, and L3) represented the downward fining sediments in the Upper Glacial aquifer (UGA). The fourth layer (L4) represented a potentially semi-confining unit (PSU), an ensemble the Gardiners Clay and the Monmouth Greensand. Layer 4 was horizontally halved into northern section (representing the UGA) and southern section (representing the PSU); the latter was sub-divided into two conductivity zones (Zone 1 and Zone 2) representing a southerly decrease in permeability in the PSU. The bottom layer (L5) represented the shallow Magothy aquifer. For further details on the hydrogeology of Long Island, see Smolensky et al. (1989), Sirkin (1982), and McClymonds and Franke (1972).

Several of the model features were either fully or partially uncertain, or have been interpreted differently by different modelers. For demonstrative purposes, the model features were classified into "fixed" features and "variable" features. The fixed features were kept constant in all model variants. For example, the precipitation rate was kept fixed at 48 inches/year (122 cm/ year); this value was approximated from the regional average precipitation rate of 48.3 inches/year for 1949-2013.

Eight model features were considered as "variable" features representing the recognized uncertainties about the hydrogeology of the study area. The uncertainty in variable features was represented by either two or three select variations, or "states" of these features. Seven of the variable features represented the "epistemic uncertainty" in the model (Table 1).

Variable	Feature	State 1	State 2	State 3
V1	L1 bottom	Uniform	Variable	
V2	L2 bottom	Uniform	Variable	
V3	PSU Extent	2-zone	3-zone	
V4	Recharge	Natural	Basins	0
V5	Segments	Yes	No	
V6	$K_{h,UGA}$ (ft/d)	High	Medium	Low
	L1	300	250	200
	L2	250	200	150
	L3	200	150	100
V7	Top of PSU	Uniform	Int.	
V8*	CHD_{North}	42’	40’	38’

Table 1: Variable features and their states (*represents aleatory uncertainty; Int.=Interpolated)

V1 and V2 represented uncertainty in the vertical discretization of the downward fining UGA sediments. V3 represented uncertainty in the northern extent of the PSU; it either begins at Zone 1 (V31) or at Zone 2 (V32). V4 represented uncertainty in how the landfill affects recharge; it is either natural with no effective liner (V41), or is diverted to recharge basins adjacent to the landfill mounds (V42), or nonexistent due to a liner system that collects it for off-site treatment (V43). V5 represented uncertainty in drains’ segmentation. In V51, the drains were simulated as linearly interpolated unsegmented polylines. In V52, the steams were divided into 3 or 4 segments whose dimensions and characteristics were individually set. V6 represented the uncertainty in the conductivity (K_h) of the UGA layers using three sets of K_h values derived from

earlier conductivity studies on Long Island (Smolensky et al. 1989; Wexler 1988). V7 represented the uncertainty in the position of top surface of the PSU; it is either shown as a uniform surface (V71), or as an undulating surface defined by interpolation of geologic boring log information (V72).

V8 represented the aleatory uncertainty in the model, the value of the northern CHD boundary. It was simulated by setting three different values of the CHD boundary – 42 feet (12.8 m) to simulate "high" groundwater conditions, 40 feet (12.2 m) to simulate "median" groundwater conditions, and 38 feet (11.6 m) to simulate "low" groundwater conditions – one for each state. These values were derived from USGS potentiometric maps for Long Island for the period 1983-2010.

A total of 288 unique 3-D, finite-difference, groundwater flow simulation models were generated by combining these variable features and their states: $V_1(2)\bullet V_2(2)\bullet V_3(2)\bullet V_4(3)\bullet V_5(2)\bullet V_6(3)\bullet V_7(2)$. The models were simulated using Visual MODFLOW v. 4.2 (Waterloo Hydrogeologic, Inc.) under steady-state conditions using the MODLFOW 2000 numerical engine and the PCG2 solver package. Simulations of twenty-three models abnormally terminated and could not be included in the analysis. It is expected that the model simulations follow a "logical ordering" of the simulated values, that is, the difference between the adjacent simulated values ("High" - "Median", "Median" - "Low") should always be positive. The logical ordering was found violated in 65 models and therefore these models were excluded. So, 200 model variants were finally evaluated using the proposed approach.

Calculation of the Area Metric

The Area Metric was calculated in four steps. Step 1 consisted of generation of $ECDF_{observed}$ and $ECDF_{simulated}$. $ECDF_{observed}$ was generated from three data points in the head observation records of a given well. The records were intermittent for a period from 1976 to

2013 at 133 observation wells in the study area. Thus, only three data points were used: the maximum, median, and minimum head observations. The use of three data points ensured equal number of steps (n=3) in each $ECDF_{observed}$, and that a conservative range of head behavior is included. The epistemic uncertainty associated with the head measurements was not incorporated into the $ECDF_{observed}$. This process was repeated for all wells generating 133 $ECDF_{observed}$. For this, $ECDF_{simulated}$, each model variant was simulated thrice by altering the values the lone aleatory feature V8 (the northern CHD boundary) from 42 feet (12.8 m), to 40 feet (12.2 m), and to 38 feet (11.6 m) for "high", "median", "low" groundwater conditions respectively. The three model-simulated head values were collated into $ECDF_{simulated}$ for each well.

Groundwater head fluctuation is a stochastic process indicating a complex groundwater regime is behavior over an area over time. This fluctuation is observed by measuring head over time at observation wells dispersed across the study area. To facilitate the calculation of the Area Metric in the present study, the stochastic process was disintegrated into individual stochastic variables (a.k.a. aleatory or random variables) two ways. First, the groundwater head fluctuations at each observation well was represented separately by 133 $ECDF_{observed}$. Second, the chronological ordering of the groundwater heads observations for an observation well was overridden with the order of magnitude develop monotonically increasing ECDFs.

In Step 2, the *well Area Metric* (A) values was calculated for each of the 133 wells by quantifying the area between the $ECDF_{observed}$ for a given well and its corresponding $ECDF_{simulated}$. This generated a set of 133 A values, one set for each of the 200 model variants. In Step 3, each set of 133 A values were used as input to generate a "model ECDF" ($ECDF_{model}$) for each model variant. In Step 4, each of the 200 $ECDF_{model}$ were compared to an ECDF of a "reference model" ($ECDF_{reference}$) a hypothetical model where A=0 for all 133 wells meaning a perfect overlap between the observed and the simulated data for each well. Then the area

enclosed between the $ECDF_{model}$ and the $ECDF_{reference}$ was quantified as the *model Area Metric* (A*) for all 200 model variants to generate 200 A* values. All calculations were made using an R code.

Results and Discussion

The $ECDF_{model}$ of the 200 model variants dispersed from about 0 feet to about 7 feet (Figure 2). This dispersion suggests that some models have better agreement with the observed data. However, they did not perform as well for all wells. The dispersion was prominent near p=1.0 where comparatively larger A values were seen. The smallest model Area Metric (A*) value was 1.25 feet (0.38 m), while the largest A* value was 2.92 feet (0.89 m).

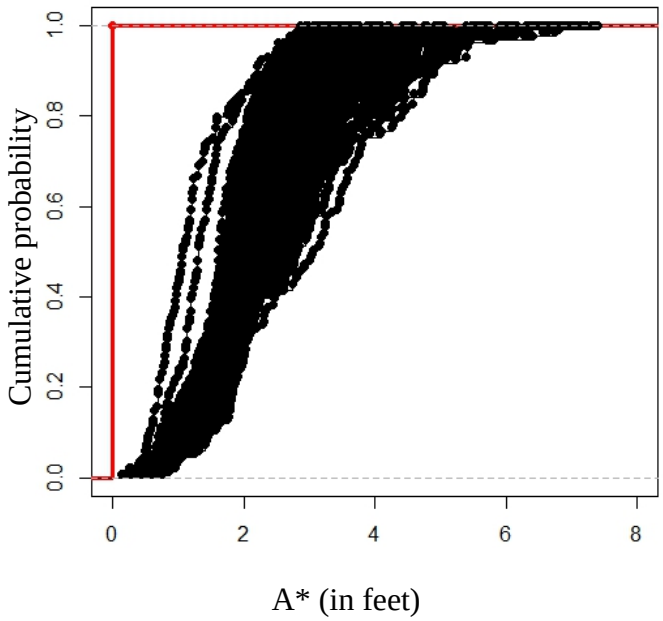

Figure 2: $ECDF_{model}$ of 200 model variants (black), along with $ECDF_{reference}$ (red)

The 133 A values of the 200 model variants were visualized as box and whiskers plots arranged in the increasing order of A* values (superimposed in red) (Figure 3). The A* values of the first 7 models were much lower, but the remaining model variants had a steady increase in A* values. The interquartile ranges increased from 0.52 feet to 2.57 feet with increasing A* values left to right. Some outliers were observed for most models.

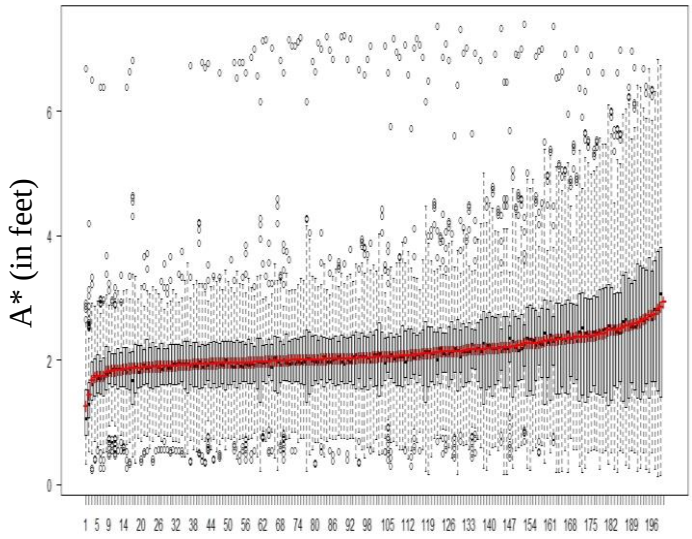

Figure 3: Boxplots of the A values for each of the 200 model variants (A* values in red)

The differences between the A* values of the states of each of the 7 epistemic variable features were analyzed by one-factor unbalanced ANOVA (Table 2). The ANOVA indicates that feature-states V12 (uniform thickness of L1), V21 (bottom of L2 close to the bottom of L1), V32 (3-zone configuration of PSU), and V61 (high conductivity setting for the UGA) resulted in model distributions with lower A* values.

Variable states	F
V11-V12	4.67*
V21-V22	9.12**
V31-V32	9.87**
V41-V42-V43	1.65
V51-V52	0.001
V61-V62-V63	101.18***
V71-V72	0.034

Table 2: Results of the 1-way unbalanced ANOVA (*= p<0.5, **=p<0.01, ***=p<0.0001)

Table 3 shows the model configurations of these top 7 models.

Model	A*	Configuration (variable states)						
178	1.25	V12	V21	V32	V43	V51	V61	V72
265	1.44	V12	V21	V31	V41	V51	V61	V72
200	1.66	V12	V22	V31	V41	V52	V63	V72
244	1.71	V12	V22	V32	V42	V52	V62	V71
177	1.73	V12	V21	V32	V43	V51	V61	V71
204	1.74	V12	V22	V31	V42	V51	V62	V72
216	1.74	V12	V22	V31	V43	V51	V62	V72

Table 3: Configurations of the top 7 models

All models contained state V12 (variable thickness of the bottom of L1). The states of V2 (bottom of L2), V3 (extent of the PSU), and V4 (recharge conditions) featured almost equally. Segmented streams (V51) was the preferred configuration in 5 of the 7 models. V63 (low permeability set for the UGA) was the least preferred feature, while V61 (high permeability set for the UGA) and V62 (medium permeability set for the UGA) appeared equally. Uniform top surface for the PSU (V71) was the preferred configuration in 5 of the 7 top models. Thus, a uniformly thick first layer, a thinner second layer, a more distinguished confining layer, and the use higher specific conductivity values may lead to better performing models.

The geospatial distribution of the means of the well Area Metric (A) showed that larger mean A values were found in wells located near the northern and the southern edge of the landfill and in the upper reaches of the streams. The A value was observed at well 96202 was distinctly high (mean A = 4.26 feet) (Figure 4).

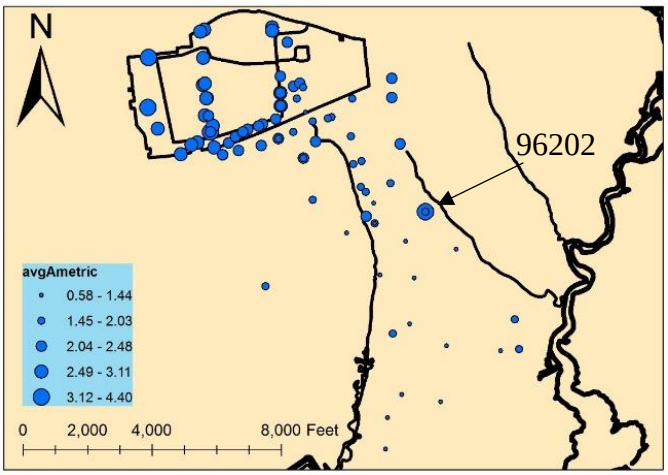

Figure 4: Geo-spatial distribution of the mean A values in the surficial plane

Three sensitivity analyses were conducted. First, the A* values were recalculated using a quartile range of head values: 1st quartile, median, and the 3rd quartile head values in Step 2 of the calculations. The difference between the quartile-based A* values and the original A* values was greater for better performing models (Figure 5). The higher ranked 179 models all had lower values, up to 0.8 feet better. However, the distribution among the top 7 models changed when quartile ranges were used. This suggested that extreme conditions controlled the relative rankings of all but the top 7 models.

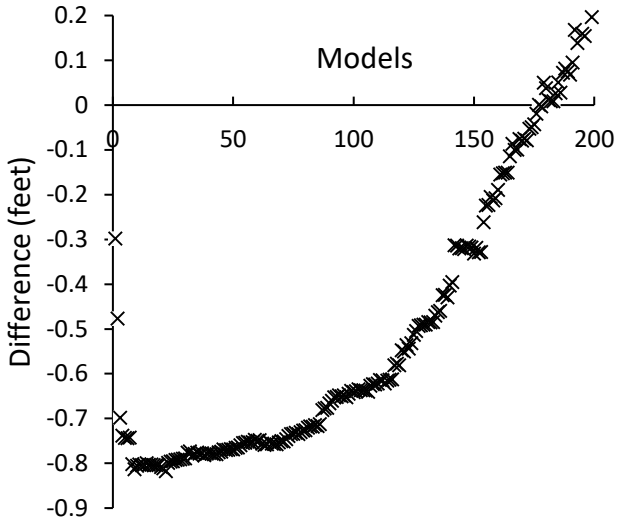

Figure 5: Difference between the A* values based on the quartile descriptors and the corresponding A* values based on the original descriptors (models arranged using original A*)

Second, the A* values for the top 7 models were recalculated excluding the outlier well S96202 that had the abnormally large mean A value. This had little effect; there was a minor change in the A* values and in the model ranks of the top 7 models (Table 4).

Model #	A*			Rank		
	Org.	New	Diff.	Org.	New	Diff.
178	1.25	1.21	0.04	1	1	0
265	1.44	1.42	0.02	2	2	0
200	1.66	1.62	0.04	3	3	0
244	1.71	1.72	-0.01	4	6	-2
177	1.73	1.73	-0.00	5	7	-2
204	1.74	1.70	0.04	6	4	2
216	1.74	1.70	0.04	7	5	2

Table 4: Change in A* and model ranks for the top 7 models with exclusion of well S96202 (Org. = Original; Diff. = Difference)

Third, the highest-ranked model, #178, was re-simulated by changing the ECDF resolution from 3 data points to 5 data points (instead of min-median-max, min-1st quartile-median-3rd quartile-max values were used). The A* value increased from 1.25 feet to 1.68 feet

and the 3-point A values were greater than the 5-point A values in 18 of 133 wells (Figure 6). It seems that adding steps increases the resolution of the ECDFs that, in turn, may increase the Area Metric values.

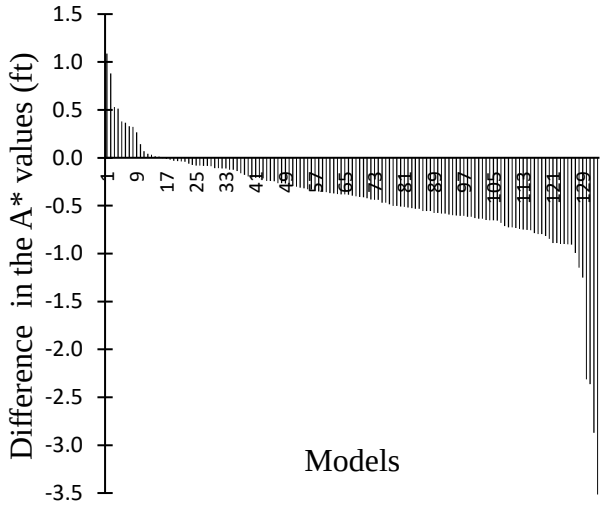

Figure 6: Differences in the 3-step and the 5-step A values for all 133 wells

Finally, the association between RMSE, a common performance indicator, and A* was analyzed (Figure 7). The RMSE values were calculated for each well for the three groundwater conditions and the resulting three RMSE values were then averaged to arrive at a model RMSE value (RMSE*) for each model variant. The correlation was positive with correlation coefficient (R^2) of 0.657. The association was apparent in poorer performing models (ranked from about 100 to 200 by A*), but not for models ranked from 1 to about 100. The RMSE gives relatively higher weights to errors of larger magnitude since the errors are squared. On the other hand, all wells are weighted equally in calculating the Area Metric. Hence, the Area Metric approach is more robust to outliers than RMSE.

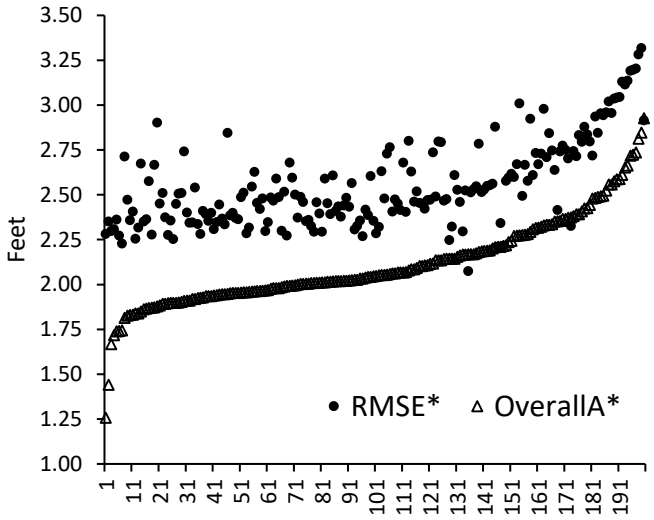

Figure 7: A* values (triangles) and their corresponding RMSE* values (circles)

Conclusion

The multi-model validity assessment using the Area Metric is firmly rooted in the pragmatic realism about how models are built and tested. Using this approach, we have shown an approach that addresses key issues in groundwater modeling.

Given uncertainty, developing and testing multiple models is a better alternative than treating a singular model as error-free. Here, the model uncertainty was explicitly represented using multiple model variants of a base landfill model.

Traditional hypothesis testing approach of binary acceptance or rejection of model's validity is not achievable given uncertainty in our understanding of real world system. Instead, here we assessed the "degree" of multiple models' validity, or the level of agreement between observed and simulated values. We did not ratify or refute the validity of any particular model, but identified models that better concerned with the observed data.

Uncertainty reduction or elimination is difficult to achieve because the potential for confounding of model errors and due to the difficulty in apportioning uncertainty to its sources in a complex and heterogeneous model. Here we offer a more pragmatic alternative of classifying

303 uncertainties into reducible (epistemic) and irreducible (aleatory) classes. Incorporation of
304 epistemic uncertainties guides future data gathering efforts. Here it showed that future data
305 collection should focus on the geology of the study area. Incorporation of the aleatory
306 uncertainty was useful in estimating the relative worth of the models under differential
307 disaggregation of the states of the groundwater flow system. For instance, the sensitivity analysis
308 showed that data with higher resolution increases our ability to distinguish among different
309 model variants.

310 Instead of calibrating a single model to obtain a better fit to observed data, a "blind
311 assessment" was conducted of the degree of multiple models' representation. This means there is
312 less chances of model over-fitting, as model configurations were not tuned or updated to obtain
313 an exact fit with the calibration data set. Instead, each of the 200 models retained their initial
314 model make-up throughout.

315 Matching model results with a singular, snapshot representation of the observed data is
316 generally thought to limit the model's applicability to other data conditions. Also, these model
317 variants were tested under a variety conditions across the whole range of system behavior
318 represented as their ECDFs of observed groundwater head data. Thus, more emphasis was given
319 to the consistency of the model behavior over the observed range of groundwater conditions and
320 not just only under median conditions.

321 The proposed approach is generalizable to other modeling studies. For example, the Area
322 Metric can be calculated for multi-dimensional observational data; for example, here, the Area
323 Metric can be calculated for each model using streamflow volumes in addition to the
324 groundwater heads and then aggregated into the model Area Metric. Also, additional procedural
325 steps can be included to accommodate pattern matching in cases of transient state models. Model

solution obtained for the inverse groundwater flow simulation studies can be used to find solution for the forward (predictive), contaminant fate and transport studies.

The proposed approach is best utilized with realistic understanding of its applicability. The model set used here as well as the uncertainties acknowledged are not exhaustive but they represent a sample very large model and uncertainty spaces. Changes in the sample model set or the acknowledged uncertainties will be reflected in model rankings. Additional assessment will be needed to expand the scope of validation beyond replicative validation to other types such as conceptual or predictive validation. The Area Metric is a descriptive measure of model's validity and it is the purview of the model user to decide if this validity is adequate for the purposes of the modeling exercise. The proposed approach is not a substitute to good modeling practices, a sustained stakeholder involvement, and to maintaining a critical distance between the modeler and the model.

These and other features of the proposed approach can increase the confidence about the representativeness of a model. A model vetted by the multi-model validity assessment using the Area Metric approach could reduce the model builder's risk of rejecting a valid model as well as the model user's risk of failing to reject an invalid model. Either ways this makes models better decision-support tools and the decisions supported by these model better informed.

Acknowledgements

The authors would like to thank Ed Hubbard, Commissioner, Division of Waste Management, Town of Brookhaven, for his support. We would like to thank Scott Ferson and Kamazima Lewiza for their persistent and scholarly guidance.

References

1. Beven, K. 2012. Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience* 344, no. 2: 77-88.

2. Beven, K. and A. Binley. 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes* 6, no.3: 279-298.
3. Bredehoeft, J. 2003. From models to performance assessment: the conceptual problem, *Ground Water* 41, no.5: 571-577.
4. 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* 197, no. 29: 2408-2430.
5. Ferson, Scott, and W. T. Tucker. 2003. Reliability of Risk Analyses for Contaminated Groundwater. *Groundwater quality modeling and management under uncertainty*. ASCE, 2003.
6. 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.
7. McClymonds, N.E. and O.L. Franke. 1972. Water transmitting properties of aquifers on Long Island, New York. *USGS Professional Paper* 627-E: E1-E24.
8. Morgan, M.G. and M. Henrion. 2006. Uncertainty: a guide to dealing with uncertainty. *Quantitative Risk and Policy Analysis*. Cambridge University Press, NY: 74.
9. Oberkampf, W. L., and M. F. Barone. 2006. Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics* 217, no. 1: 5-36.
10. Oreskes N. 1998. Evaluation (not validation) of quantitative models. *Environmental Health Perspectives* 106, no. 6: 1453-1460.
11. Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 263, no. 5147: 641-646.

12. Peterson, D.S. 1987. Groundwater recharge in Nassau and Suffolk counties, New York. *U.S. Geological Survey Water Resources Investigations Report* 86-4181: 19.

13. Poeter, E., and D. Anderson. 2005. Multimodel ranking and ground water modeling. *Ground Water* 43, no. 4: 597-605.

14. Refsgaard, J. C., S. Christensen, T.O. Sonnenborg, D. Seifert, A.L. Højberg, and L. Troldborg. 2012. Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources* 36: 36-50.

15. Romanowicz, R. and R. MacDonald. 2005. Modeling uncertainty and variability in environmental systems. *Acta Geophysica Polonica*, 53(4): 401-417.

16. Roy, C. J. and W.L. Oberkampf. 2011. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computational Methods in Applied Mechanics and Engineering* 200, no. 25: 2131-2144.

17. Singh, A., S. Mishra, G. Ruskauff. 2010. Model averaging techniques for quantifying conceptual model uncertainty. *Groundwater* 48, no. 5: 701-715.

18. Sirkin, L.A. 1982. Wisconsinan glaciation of Long Island, New York, to Block Island, Rhode Island. In *Late Wisconsinan Glaciation of New England*, B. Stone and G. Larson (Ed), Kendall/Hunt: 35-59.

19. Smolensky, D.A., H.T. Buxton, and P.K. Shernoff. 1989. Hydrologic framework of Long Island, New York. *USGS Hydrologic Investigations Atlas* HA-709: 3 sheets, Scale 1:250,000.

20. Waganer, T. and H. Gupta. 2005. Model identification for hydrological forecasting under uncertainty. *Stochastic Environmental Research and Risk Assessment* 19, no. 6: 378-387.

21. Wexler, E.J. 1988. Ground-water flow and solute transport at a municipal landfill site on Long Island, New York, part 1, hydrogeology and water quality. *USGS Water Resources Investigations Report* 86-4070: 53 p.

22. Yapo, P., H. Gupta, and S. Sorooshian. 1998. Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, 204: 83-97.