Stony Brook University

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

A Dissertation Presented

by

Omkar Aphale

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Technology, Policy, and Innovation

Stony Brook University

December 2015

Stony Brook University
The Graduate School

Omkar Aphale

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

David J. Tonjes
David J. Tonjes – Dissertation Advisor
Research Professor, Dept. of Technology and Society

David Ferguson
David Ferguson - Chairperson of Defense
Professor, Dept. of Technology and Society

Scott Ferson
Scott Ferson – Third Inside Member
Adjunct Faculty, Dept. of Technology and Society

Henry Bokuniewicz
Henry Bokuniewicz – Outside Member
Professor, School of Marine and Atmospheric Sciences

Kamazima Lewiza
Kamazima Lewiza - Outside member.
Associate Professor, School of Marine and Atmospheric Sciences

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

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

by

Omkar Aphale

Doctor of Philosophy

in

Technology, Policy, and Innovation

Stony Brook University

2015

A model's validity, or its goodness-of-fit to the real world system, is commonly assessed by quantifying the level of agreement between the observed data and their corresponding model-simulated outputs. However, the observed data could be uncertain given inaccuracies in the observational tools and techniques while the model-simulated outputs may be incomparable since models are simplified versions, and not exact replicas, of the real world system. This limits the abilities of the traditional validation approaches.

Here, an alternative approach called the area metric (Ferson et al. 2008) was adopted for multi-model validation assessment. This approach quantifies the level of agreement between the observed data and the model-simulated outputs expressed as their respective empirical cumulative distribution functions.

The area metric approach was used to assess the validity of multiple model variants of a base model that simulates the groundwater conditions in the vicinity of the municipal landfill in the Town of Brookhaven, NY. Uncertainties regarding the configuration and the characteristics of a groundwater system were represented by developing 288 model variants of varying conceptualizations of the base model. These models' validity was assessed over a conservative range of groundwater head data from 133 observation wells. Based on the calculated model area

metric values, the models were ranked and the 10 models with the lowest area metric values were selected as conforming best to the data.

In this way, the area metric-based multi-model assessment selects, from a model space, better representations of groundwater flow systems. It avoids overfitting a single model to a particular system state and facilitates incorporation of the epistemic and aleatory uncertainties into the validation process. In addition, the approach acknowledges that finding an exact correspondence between observed data and simulated output is difficult, given all aspects of model uncertainty. Therefore, the area metric-based multi-model validation approach explicitly represented model uncertainty using multiple model variants and the degree these models replicated real conditions was tested over a range of observed data.

Overview of the dissertation

The dissertation is divided into six chapters. Chapter 1 contains the general introduction to the topic and includes discussion on (i) modeling and its application in groundwater science, with an illustrative example; (ii) model uncertainty and its various classifications; (iii) model validation, and on (iv) multi-model analysis (MMA) with four select approaches. Chapter 2 delineates the objectives and the scope of this study. Chapter 3 describes the method; the details of (i) the study area, (ii) the model used, (iii) the fixed and variable features and their states, and (iv) the theory and calculation of the area metric are discussed here. Chapter 4 describes the results specific to the case of the application of validation metric for the multi-model analysis. Chapter 5 includes general discussion about the broader theoretical and methodological implications of the proposed approach. Conclusions are made in Chapter 6. References are included thereafter. Appropriate additional information is attached to the report in the form of three appendices at the end of the report.

Dedication Page

To Aaī and Papa

Table of Contents

List of Figures/Tables/Illustrations

List of Figures

List of Tables

Acknowledgments

I am truly fortunate to have had the opportunity to work with to my advisor, Dr. David J. Tonjes. I thank him for the persistent guidance and mentorship he provided to me both in academic and personal matters, from the day I joined the program through to completion of this degree. Dr. Tonjes has shown tremendous patience while working with me and I wish I could emulate his genuinely good nature and down-to earth humility.

I would also like to thank my committee members, Dr. Henry Bokeniewicz, Dr. Scott Ferson, and Dr. Kamazima Lewiza for their expert guidance, thought provoking suggestions, and the accommodating nature that each of them offered to me. In a similar vein, I'd like to recognize Dr. Gilbert Hanson, Dr. Robert Cerrato, Dr. R.L. Swanson, Dr. Sultan Hameed, Dr. Ed Kaplan, and Dr. Sheldon Reaven for the sagacious contributions that each of them made to my intellectual growth during my years of study at the Stony Brook University.

I am very grateful to Dr. David Ferguson, Chair, Department of Technology and Society. Dr. Ferguson was always there when I needed him, advising and encouraging me to stay on course. I owe an immense gratitude to the staff at the Department, especially to Rita, Marypat, Joyce, and to Romayne for their relentless emotional support and motherly affection.

This research was supported by the Department of Solid Waste Management, Town of Brookhaven, NY. I am grateful to have support of Ed Hubbard, Commissioner of Waste Management, Town of Brookhaven, NY. I learned many useful things outside of school in the company of Micheal DesGains (Town of Brookhaven), Russell Wetjen (Mclean Associates), Anthony Caniano (Dvirka and Bartilucci, Inc.), and Paul Misut and others (USGS Water Science Center, Coram). I am grateful for their collegiality and friendly guidance.

I thank my doctoral colleagues Krista Greene-Thyberg and Lori Clark helped me on a number of academic projects and shared their time and thoughts with me for all these years. Also, I thank my friends at Stony Brook University – Saurab Joglekar, Akshay Patil, Ravi Dey, Srikanth Mallikarjun, and Prahlad Deshpande – for all the good times we had.

I am forever indebted to my parents, whose unconditional love and tireless support has inspired me to pursue a higher purpose in life. I thank my sister Aasawari and my brother-in-law Pushkar for their immovable faith on me. I thank my wife Sonal whose companionship has rekindled my hopes for a better tomorrow.

Lastly, I thank God to whom I surrender everything.

Chapter 1

Introduction

Overview:

This chapter contains the general introduction to various basic concepts fundamental to this dissertation. Firstly, concepts of models and their applications in groundwater science are introduced. An illustrative example of this application, of a flow simulation and contaminant transport model for groundwater near the municipal landfill in the Town of Brookhaven in New York, is briefly explained. Secondly, the chapter discusses model uncertainty, its definition, its classifications based on the class of uncertainty (conceptual, input, and parameter), the levels of uncertainty (from complete determinism to total ignorance), and the nature of uncertainty (aleatory or epistemic). Thirdly, the chapter focuses on the concept and the need to develop multiple models given model uncertainty. Fourthly, the concept of model validation and its multiple categories are briefly described; focus is placed on replicative validation. This leads to the discussion of the concept of multi-model analysis (MMA). Here, four select approaches to the MMA – model selection based on the Akaike Information Criterion (AIC), multi-model averaging, multi-objective optimization (MOO), and Generalized Likelihood Uncertainty Estimation (GLUE) – are discussed. Subsequently, two limitations of these MMA approaches – the epistemic and aleatory uncertainty associated with the observed data, and the simplified representation of the real-world system – are highlighted. Lastly, it is suggested that the traditional approaches should be revised and new validation tools that incorporate model uncertainties should be developed.

1.1. Simulation Modeling

Simulation modeling is an important decision-support tool that enables better-informed decision making and therefore it is widely used by policy-makers, administrators, engineers, and scientists and researchers working in different scientific disciplines (Fetter 2001, p. 224). A model is "any device or tool that is developed as a simplified representation, or an approximation of the real world field situation" (Anderson and Woessner 1992, p. 2). Models can be descriptive and predictive. Descriptive models help to collate, organize, and store the experimental observations and empirical data about real world system. Therefore, these models offer a powerful, low-cost learning environment to test theories that enhance our understanding of real world systems (Bredehoeft 2005, Konikow and Bredehoeft 1992). Models with predictive capabilities allow the decision makers to simulate future behavior of a system and evaluate the outcomes, risks, and payoffs of their decisions. Simulation modeling can be used for "What If?" analysis where actual changes may be costly, unethical, or produce irreversible effects (Sterman 2006; Lahsen 2005). Different scenarios and alternate environments can be simulated, normal conditions can be disrupted, dramatic stresses can be introduced, and effects of decisions can be evaluated at comparatively nominal costs (Fetter 2001, p. 96; Anderson and Woessner 1992, p.3). Simulation modeling is described as the "third branch of science" because of its variety of applications, generalizability, and widespread use, after theory development and experimentation (Sterman 2006).

1.2. Modeling in Groundwater Science

The present exercise focuses on the use of simulation modeling in hydrogeology, the discipline in science that studies the "interrelationships of geologic materials and processes with water" (Fetter 2001, p. 3). Two primary uses of hydrogeologic models are the simulation of groundwater flow simulation and the simulation of the solute transport (Konikow 1996); specifically, the focus is on hydrogeologic models that simulate groundwater heads ("groundwater flow models" or "groundwater models").

The application of groundwater flow models can be interpretive if the modeling objectives are (i) to characterize the hydrogeologic regime of the study area, (ii) to augment conceptual understanding of its functioning, and (iii) to organize the different types of data collected about the hydrologic system. On the other hand, application of groundwater flow models is said to be predictive if the modeling objective is to describe the future behavior of the system, either in its response to changes in normal conditions, to extreme or sudden stresses, or to various remedial measures (Anderson and Woessner 1994, p. 4-5).

Models can be classified based on the type of medium in which the groundwater flow is to be simulated. Some models simulate groundwater flow in unconsolidated sediment deposits, such as models developed in Long Island in New York (Buxton and Smolensky 1999; Gureghian et al. 1980; Harbaugh and Getzen 1977; Pinder 1973). Other models simulate flow in the fractured media or karst systems where the consolidated deposits, such as rocks, have low permeability in general but they can be highly conductive in areas with interconnected pores, fractures, and fissures (Worthington 1999; Palmer and Palmer 1999, p. 272). Groundwater models are categorized according to modeling techniques: physical scale models, analog models, and mathematical models (Fetter 2001, p. 515; Wang and Anderson 1982, p. 2).

A physical scale model is analogous to a miniature architectural model. These models are typically "sand tank" models that are developed using materials that have characteristics of the sediments found in the hydrogeologic system. These could be working models that demonstrate the functioning of the aquifer systems so that water is added into the model to simulate recharge, or a dye is added to demonstrate the flow of contaminants. Physical scale models are used typically as non-technical demonstrations to general audiences.

An analog model simulates the groundwater flow and aquifer conditions using electrical circuits. The flow of groundwater is considered analogous to the flow of electricity. For

example, the change in the groundwater heads due to hydraulic gradient is considered analogous to the change in the voltage in an electrical model. Electrical flow passes through capacitors and resisters representing the varying aquifers' storativity and transmissivity. In the 1960s, the United States Geological Survey (USGS) generated numerous case-specific analog models examples in its Phoenix laboratory. The components of the analog models were organized onto a pegboard in a rectilinear grid format and the outputs were measured using voltmeters (Bredehoeft 2012).

A groundwater flow problem can be solved mathematically when the initial and the boundary conditions are known and when the functional relationship between the aquifer and fluid properties is expressed in the form of mathematical equations. Mathematical models are of two types: analytical models and numerical models (Fetter 2001, p. 516; Anderson and Woessner 1992, p. 2).

An analytical model is built on simple assumptions and the flow equations are solved using calculus. A famous example of an analytical model is the Darcy's Law where French hydrogeologist Henry Darcy described the relationship between the discharge of water and hydraulic head (Darcy 1856) (Equation 1.1):

$$Q = -KA\left(\frac{h_1 - h_2}{l_1 - l_2}\right) \tag{1.1}$$

where,

Q = discharge (L^3T^{-1}),

K = hydraulic conductivity (L^1T^{-1}),

$h_1 - h_2$ = change in head between two points 1 and 2 in the porous medium in a pipe (L^1),

$l_1 - l_2$= distance between the points (L^1),

A = Cross-sectional area of the pipe (L^2)

A numerical model can be used if the assumptions in the analytical solutions are considered too simplistic to represent the complex settings of the real world groundwater regime. Here, the governing flow equations are solved by discretizing the spatio-temporal continuity of the equations and by approximating the aquifer and fluid properties within the discretized model domain (Konikow 1996). For example, Laplace's equation is the governing equation for steady state groundwater flow via a three-dimensional, isotropic, homogeneous aquifer (Equation 1.2):

$$\frac{\partial}{\partial x}\left(-K\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(-K\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(-K\frac{\partial h}{\partial z}\right)=0 \quad (1.2)$$

where,

K = hydraulic conductivity (L^1T^{-1}),

∂h = partial derivative of hydraulic head

∂x = partial derivative of space in X direction,

∂y = partial derivative of space in Y direction

∂z = partial derivative of space in Z direction

Numerical solutions are flexible in their assumptions and they are better at incorporating the complexities of heterogeneous, physically distributed hydrogeologic systems compared to analytical solutions. Therefore, numerical solutions are more commonly used than analytical solutions (Anderson and Woessner 1992, p. 20). Numerical solutions are classified either as "finite difference" or "finite element" solutions (Wang and Anderson 1982, p. 3). A finite difference grid divides the continuous domain into a mesh of rectilinear cells. As the dimensions of the cell decreases, the mesh begins to resemble the continuous surface of the domain more closely (Konikow 1996).

The finite difference grid is sub-classified into two groups: "block centered" grids (Figure 1.1-a) and "mesh centered" grids (Figure 1.1-b) based on the position of the node points where the equations are solved to obtain simulated outputs. The nodes are located at the center of grid cell in a block centered grids, while the nodes are located at the corners of the cell in a mesh centered grid. In case of finite element grid, the nodes are located on the edges of the element and the grid is not rectilinear. The most common format for a finite element grid is a triangular element grid (Figure 1.1-c).

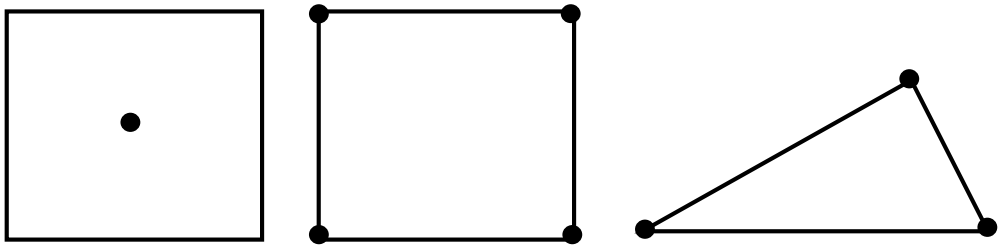


Figure 1.1: Grid cell of a (a) finite difference block centered grid, (b) finite difference mesh centered grid, and (c) finite element grid. The solid circles represent nodes.

1.3. Standard Example of Groundwater Modeling

The following is a representative example of the application of simulation modeling for groundwater flow and solute transport.

The Town of Brookhaven landfill is located in the hamlet of Brookhaven, Suffolk County, New York (Figure 1.2). The landfill, constructed in 1972, was one of the first artificially lined landfills in the country, but the liner system failed sometime after installation causing widespread groundwater contamination in the direction of groundwater flow (Dvirka and Bartilucci 2010). The impact is mainly on Upper Glacial aquifer, the water table aquifer. The USGS entered into a cooperative agreement with the Town to investigate the groundwater contamination. The work conducted under this agreement was documented in a series of reports (Wexler 1988a; Wexler 1988b; Wexler and Maus 1988).

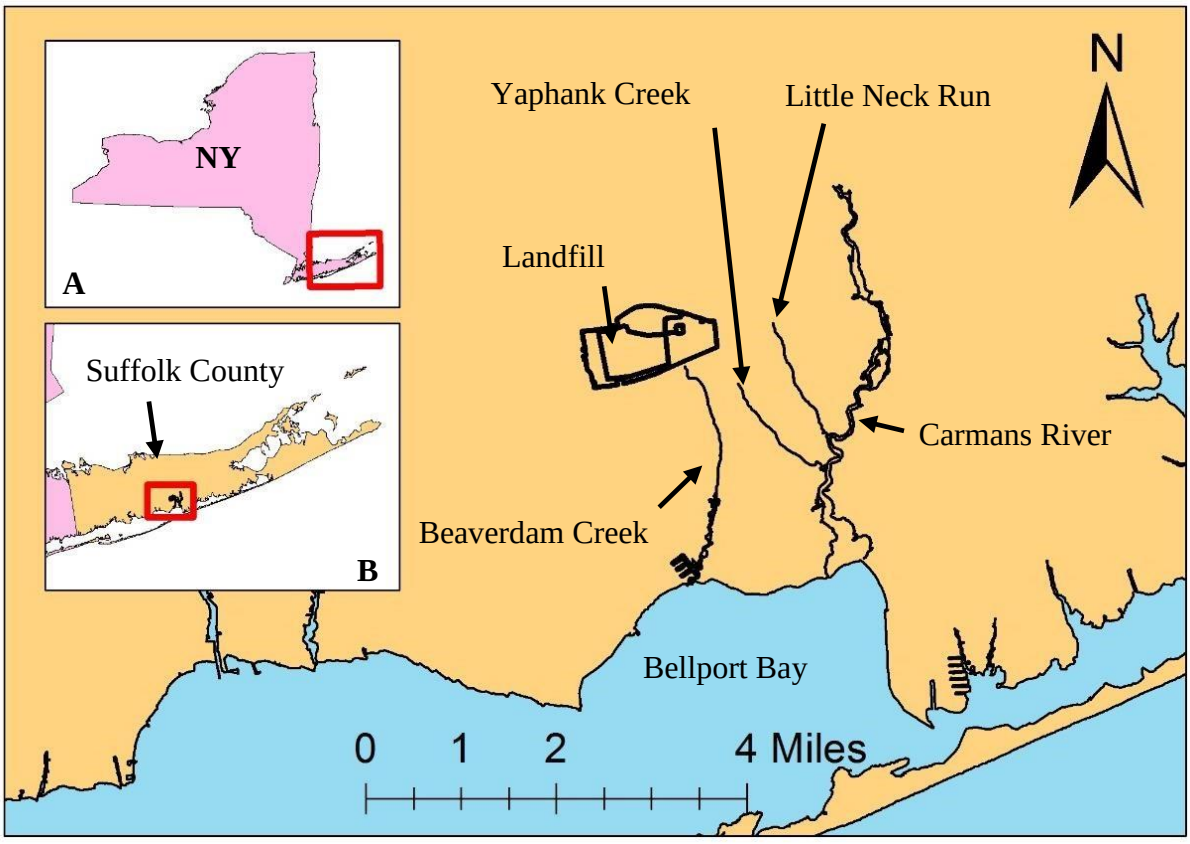


Figure 1.2: Town of Brookhaven West Management Facility, New York state (inset A), and Suffolk County, NY (inset B) (red square indicates area of detail)

The hydrogeology and existing water quality in the vicinity of the Brookhaven landfill site was described in the first report (Wexler 1988a). The second report discussed the two dimensional, finite element, steady state model simulating the groundwater flow in the Upper Glacial aquifer (Wexler and Maus 1988). The simulated groundwater heads were used to develop contours of equipotential lines depicting the height and the pattern of movement of groundwater in the Upper Glacial aquifer (Figure 1.3).

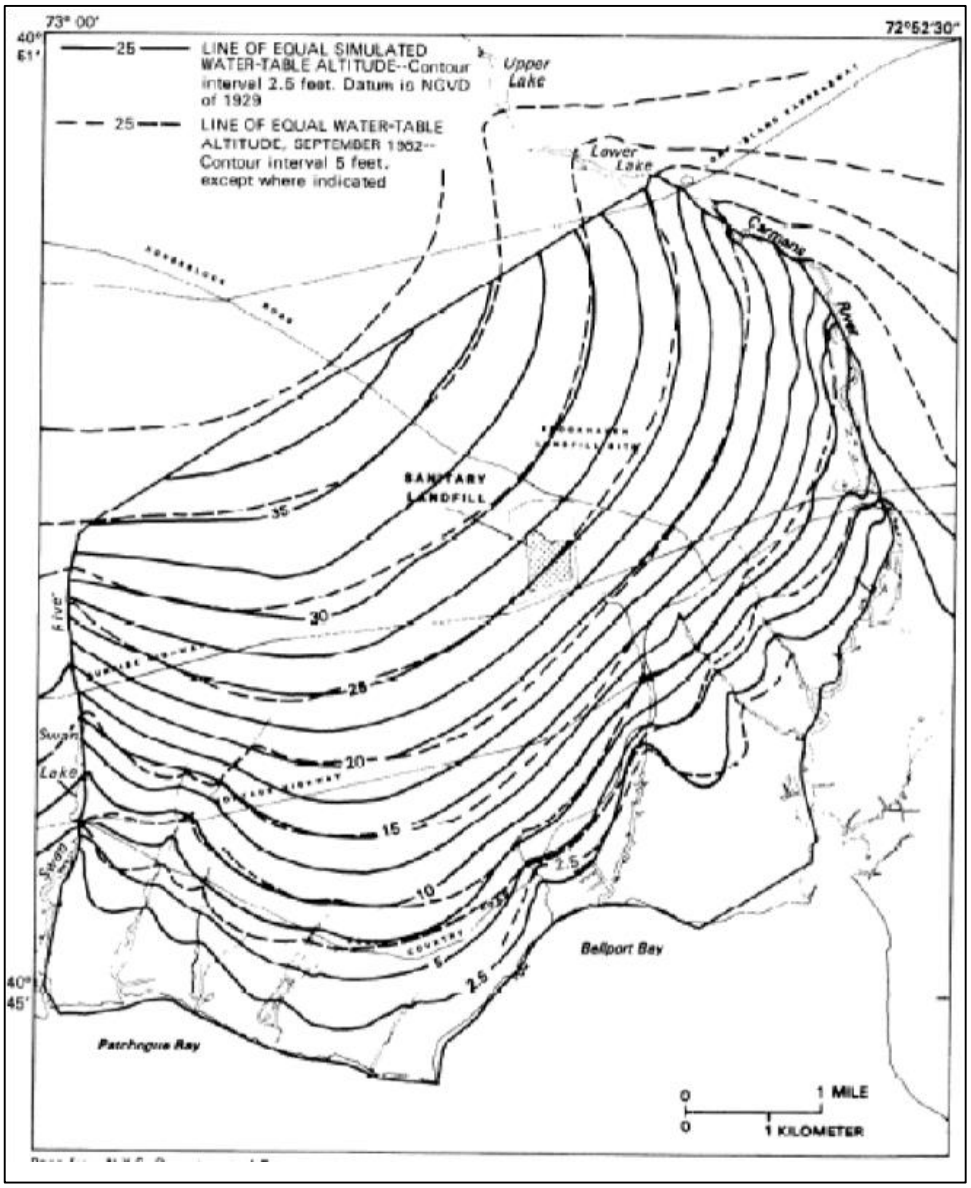


Figure 1.3: The equipotential lines of the simulated heads (solid lines) and that of the observed heads (dashed lines) (Wexler and Maus 1988)

In addition, a schematic diagram of the water budget was prepared that indicated the rate and the direction of simulated flows across different model components (Figure 1.4). Also, flowlines were generated to depict the velocity of the groundwater flow (Figure 1.5-a), as well as the movement of groundwater from the landfill site (Figure 1.5-b).

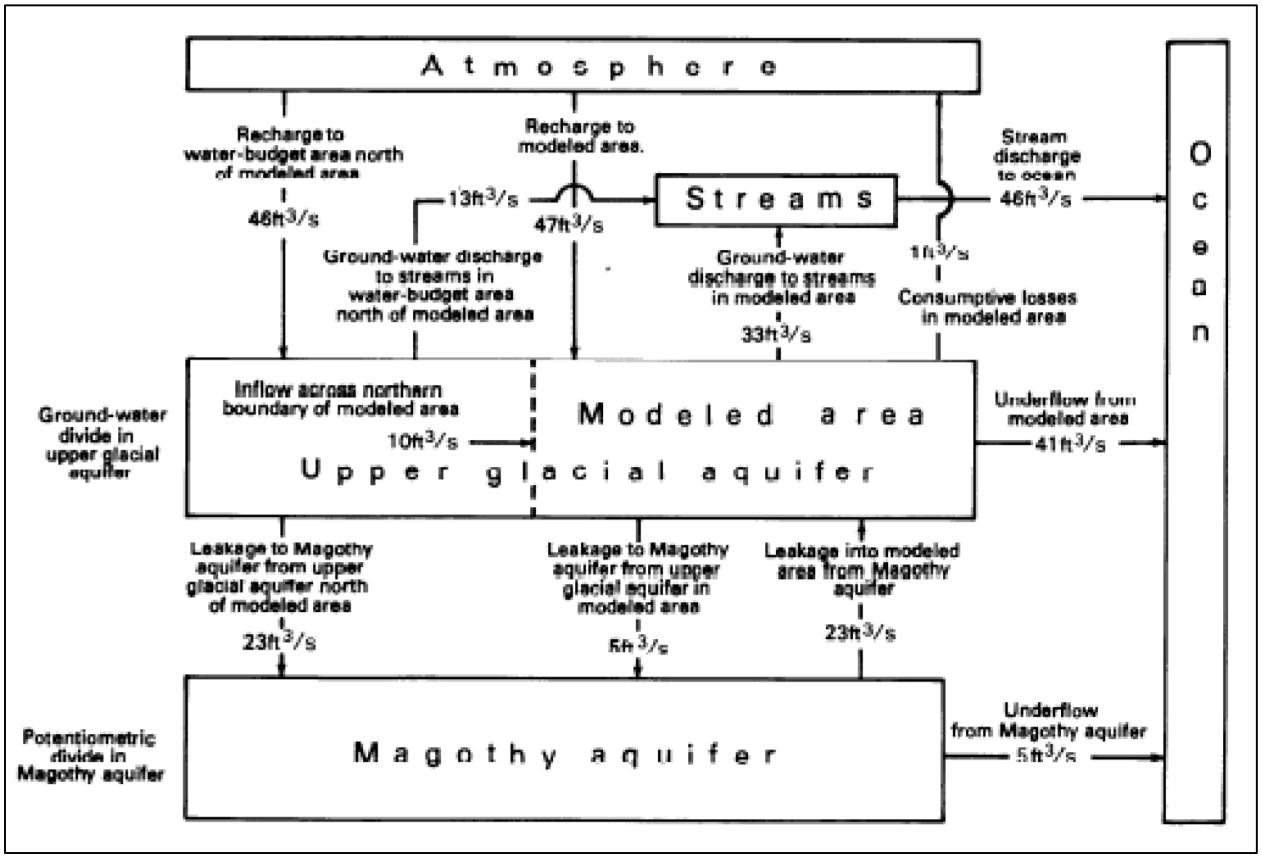


Figure 1.4: Schematic diagram of water budget (Wexler and Maus 1998)

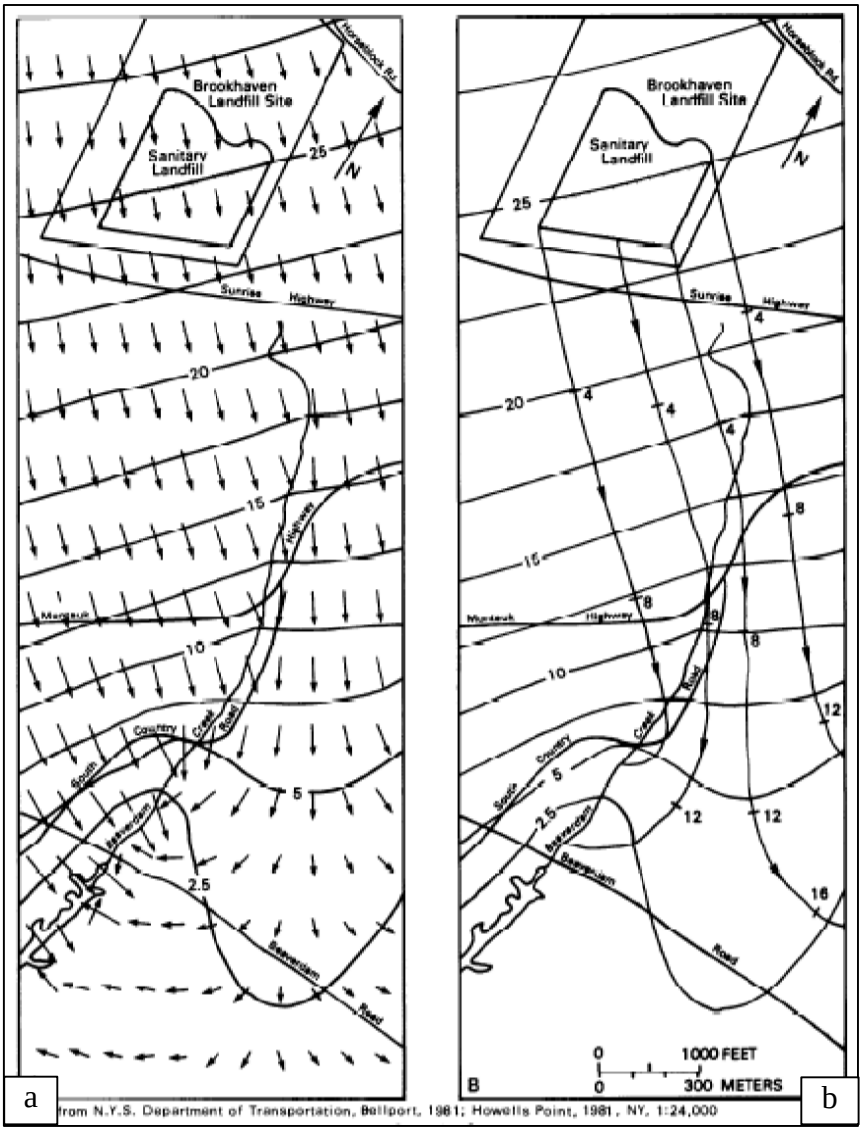


Figure 1.5: Flow lines showing (a) velocity, and (b) advective movement of the groundwater flow at 4 year intervals (Wexler and Maus 1988)

The model developed in the second report formed the basis for the construction and simulation of a two-dimensional, transient state solute transport model simulated for a period of 12 years using the SUTRA code. This model simulated the advective-dispersive migration of chloride, a conservative chemical species indicative of landfill leachate. Contours of chloride concentrations were developed at two-year intervals (Figure 1.6-a-c).

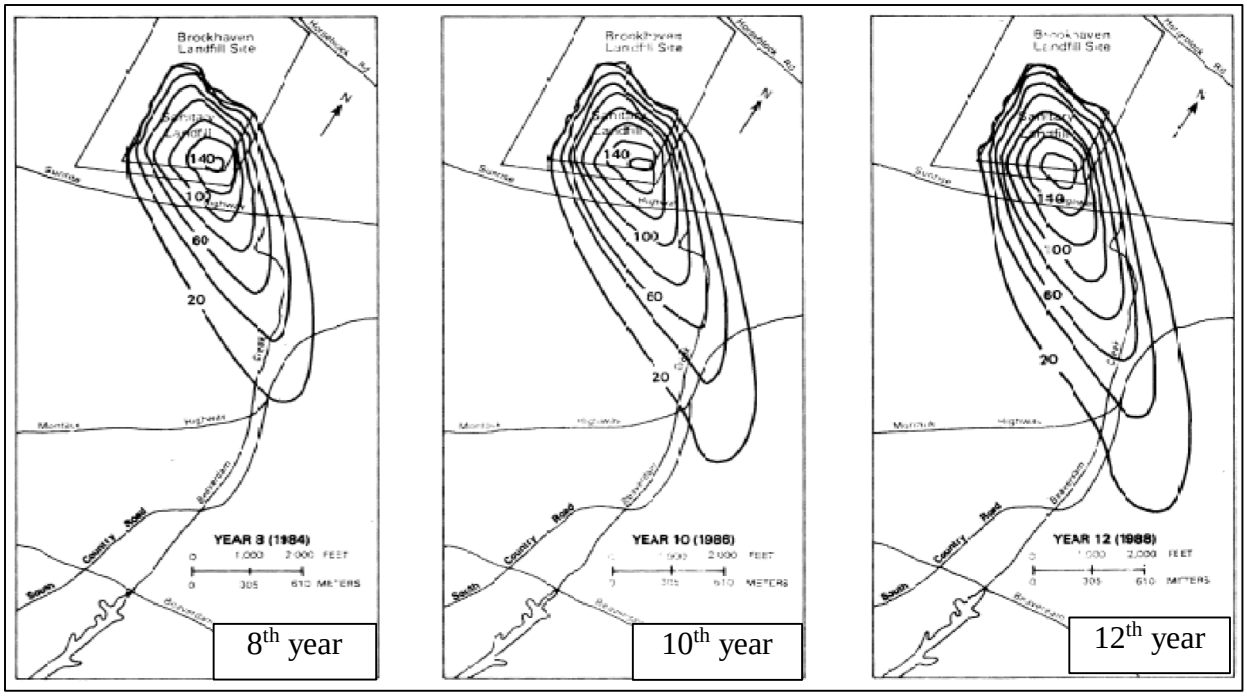


Figure 1.6: Contours for the chloride concentrations for 8th, 10th, and 12th simulation year (Wexler 1988b)

The contaminant transport model was used to simulate the effects of remediation strategies on the plume advancement. For example, Figure 1.7-a-c show contours of chloride concentrations as the effect of pumping of the contaminated groundwater through four recovery wells installed adjacent to the headwaters of the Beaverdam Creek. Contours indicated progressive containment of the plume and declines in concentrations from the 8th to the 12th year. Sensitivity analyses were carried out to evaluate the impacts of changing model parameters such as the effective porosity and the longitudinal and transverse dispersivity.

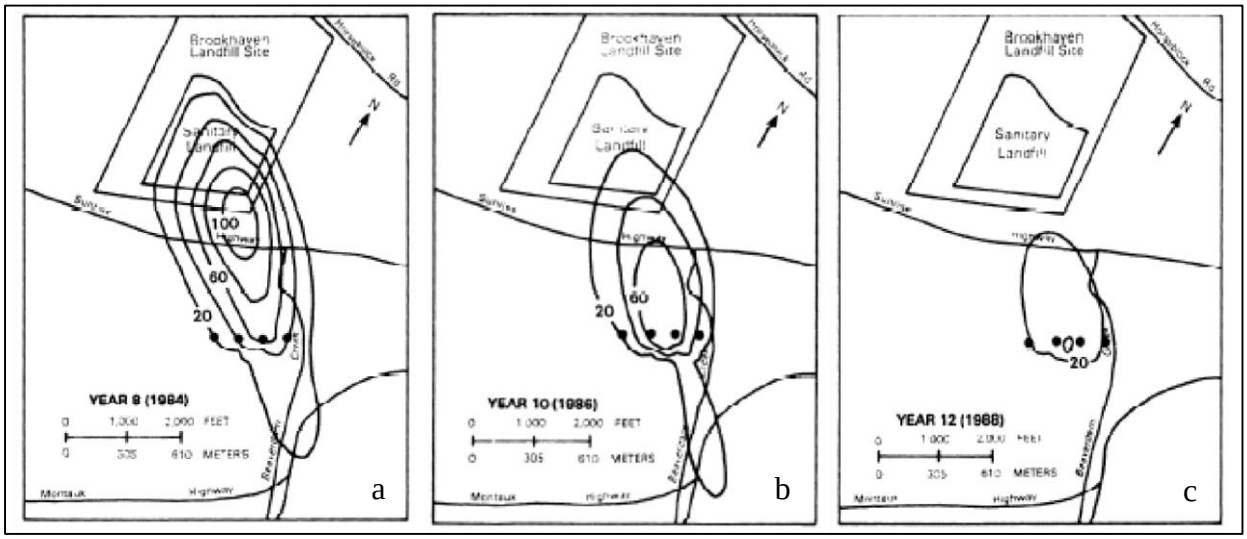


Figure 1.7: Contours for the chloride concentrations for the (a) 8th, (b) 10th, and (c) 12th simulation year the groundwater is pumped through a line of recovery wells (Wexler 1988b)

This series of USGS reports was the basis for the Town's decision to cap the landfill. Capping of the landfill cells would prevent direct infiltration of precipitation into the waste mass that, in turn, would prevent generation of leachate. In addition, a long term water quality monitoring program was initiated; newer monitoring wells were developed in the vicinity of the landfill, particularly in the downgradient area (David Tonjes, personal communication; Dvirka and Bartilucci 2010).

Both the groundwater flow and solute transport models were simplified versions of the real world hydrogeologic settings and contaminant plume conditions at the landfill site. For example, the impact of leachate was mainly on Upper Glacial aquifer and therefore both models were two-dimensional. Also, the saltwater-freshwater boundary was considered stationary. In addition, aggregate approximations, or parameters, were used to represent heterogeneous model inputs. For example, fixed parameter values were used as inputs; the annual precipitation rate was fixed at 47.4 inches/year and the recharge to the Upper Glacial aquifer was fixed at 24.6 inches/year or 52 percent of the annual precipitation. In the solute transport model, the leachate discharge was simulated through a series of discharge nodes located inside the perimeter of the landfill at a rate of 24.6 inches/year at an average source concentration of 875 mg/L. Wexler and Maus note that the interpretation of the model results "should always include consideration of the two-dimensional approach and other simplifying assumptions made in the model development".

The goodness-of-fit of the models to the real world conditions was achieved by calibrating the models with respect to observed data obtained during a singular observation/sampling period from different locations in the model domain. The groundwater flow model was calibrated using a singular set of head observations made during September 1982 at 93 wells screened at different levels in the Upper Glacial aquifer. The solute transport model was calibrated using the chloride concentration data obtained during October-December 1982. Initial input values of parameters were adjusted via trial-and-error process until a reasonable match between the simulated and the observed data used for calibration was obtained. The least-squares method was used to determine the best fit during the calibration procedure. Qualitative verification was used to determine the goodness-of-fit of the models to other aspects of the model behavior; for instance, verification of the model was limited to the visual comparison of equipotential contours developed from measurements and the simulation.

Wexler and Maus acknowledged the uncertainty associated with the model development and said that the accuracy of the model could be improved with increased knowledge of the hydrogeology of the study area. Also, they acknowledged that the parameterization of the heterogeneous characteristics of the study area, particularly of the hydraulic conductivity, may have underrepresented the distributed nature of these characteristics. Additional hydrogeologic testing and more measurements were recommended for improved model performance.

1.4. Modeling Uncertainty

A key challenge in modeling of groundwater flow systems is to deal with uncertainty associated with the configuration of these systems. Uncertainty is any deviation from the ideal of complete deterministic knowledge of the relevant system (Walker et al. 2003). Uncertainty implies that "in certain situations a person does not possess information which is quantitatively and qualitatively appropriate to describe, prescribe or predict deterministically and numerically a system, its behavior or other characteristics" (Zimmermann 2000). Model uncertainty is not singular, but rather it is a collective noun used to denote a multitude of model uncertainties. It is associated with the conceptual and mathematical structure of the model as well as the parameter values that are entered into a model (as characterized by its structure) (Neuman 2003).

Model uncertainty can be classified under three different classification systems (Walker et al. 2003). These classification systems are based:

- in terms of the location of uncertainty (section 1.4.1)
    - conceptual
    - input
    - parameter
- in terms of the level of uncertainty (section 1.4.2),
    - complete determinism
    - statistical uncertainty
    - scenario uncertainty
    - recognized ignorance
    - total ignorance
- in terms of the nature of uncertainty (section 1.4.3).
    - aleatory
    - epistemic

Each classification system and the sub-classes that fall under each classification system are described in detail below.

1.4.1. Location of Uncertainty

Uncertainty manifests in different locations of the modeling process; the key classes are: conceptual uncertainty, input uncertainty, and parameter uncertainty (Beven 2012; Refsgaard et al. 2006).

1.4.1.1. Conceptual Uncertainty

Conceptual uncertainty is the uncertainty that arises from the modeler's subjective understanding of the real world system (Refsgaard et al. 2006). Conceptual uncertainty is also known as model structural uncertainty because it indicates inadequate representation of the subsurface geologic framework, the hydrogeologic interactions, and the dynamics of the physical process (Walker et al. 2003). It also indicates the model's inability to properly explain all of the available observations of input variables and parameters (Singh et al. 2010). Conceptual uncertainty also arises when there is uncertainty regarding the objective of the modeling exercise (Romanowicz and MacDonald 2005). Bias introduced into the model by conceptual uncertainty may exceed the bias introduced by the uncertainty associated with the observational data and the parameter values (Reilly and Harbaugh 2004).

Since the decision to include or exclude real world system component(s) is subjective, models reflect the modeler's comprehension of the real world system. Although empirical data are objective, subjective beliefs of the modeler may favor one piece of data over another and this choice may not be well-grounded (Zimmerman 2000).

For the example of the Brookhaven model, conceptual uncertainty occurs because some investigations believed Gardiners Clay to be missing or to lie well south of the landfill site. The standard conceptual model of the Long Island aquifer system includes a low permeability unit, Gardiners Clay, between the Upper Glacial and Magothy aquifers. Different opinions exist about the local elevation and thickness of the Gardiners Clay underneath the landfill site and vicinity. Gerathy and Miller (1985) suggested absence of the Gardiners Clay unit beneath a previously proposed ashfill site that was located across the Horse Block Road in Yaphank, north of the landfill site. On the other hand, Voorhis (1986) reported presence of (i) discontiguous thin bands of brown clay with sandy facies at Patchogue-Yaphank Road, north of the landfill site, at an elevation of 137 feet to the mean sea level (msl); (ii) a 15 feet layer of sandy clay at an elevation of -89 feet msl on Bellport Station Road, west of the landfill site; and (iii) a 28 feet thick layer of

the Gardiners Clay on the Head of the Neck Road well site south of the landfill site at an elevation of -118 feet msl. The clay found at this site was interbedded with the Upper Glacial deposits in the form of thin bands. Buxton and Modica (1992) suggested that the Gardiners Clay is not present underneath the site. Dvirka and Bartilucci (1994a) suggest that the Gardiners Clay unit extends north of the Long Island Expressway that is present north of the landfill site. In other instances, DeLaguna (1963) and Weiss (1954) studied the hydrogeology of the Brookhaven National Laboratory (BNL) and vicinity and suggested that the Gardiners Clay is present up to the northern boundary of the BNL site and is contiguous with the south shore facies of the Gardiners Clay at an elevation of -90 to -130 feet msl. Conversely, Smolensky et al. (1989) restricted the extent of the Gardiners Clay slightly north of Sunrise Highway near the BNL site.

Additionally, inappropriate translation of the conceptual model into the computer code may give rise to technical uncertainty that may remain undetected and cause errors in the model output although the model conceptualization is reasonably accurate (Walker et al. 2003).

1.4.1.2. Input Uncertainty

Part of the model conceptualization process is to subjectively determine the major structure of the model. Input uncertainty reflects disagreement regarding structural details, such as the extent of aquifer features, various types of stresses, and boundary and initial conditions (Beven 2012; Morgan and Henrion 2006, p. 56; Konikow 1996). When modelers make decisions regarding the spatio-temporal boundaries of the model domain, continuity needs to be maintained between the part of the system included in the model domain and the external forces that may influence the system inside the model domain. Hence, input uncertainty refers not only to the uncertainty about the features included in the model domain, but also uncertainty about the type and the magnitude of the external forces (Walker et al. 2003).

Qualitative or inexact information adds to the input uncertainty because it requires subjective interpretation of the input definition. Table 1.1 shows excerpt from select geologic boring logs taken from five locations near the Brookhaven landfill.

Location	Description
MW11M	160'-182' : Black to silvery black micaceceous, lignitic SILT, some clay, little to some fine sand 185'-187' : No recovery-drilling indicates clay 190'-192': Same As Above 195': Drilling change at approximately indicating sand
MW4-D	172'-174': 0-1.2'→ Br f S, a(+)$; no prt, mica, dense, wet 1.2-2.1'→Gy br C l(+), fs, tr(-) c: ang qtz, faint br prt, damp, stiff, mica 2.1-2.2'→ Or br f(-) c S, l(-) $yc, tr (-)f(+)cg; rnd, stiff, damp
103140	Medium to coarse dark brown sand-some gravel and mica found throughout
S72813M	180-187: Gardiners Clay (sandy facies, some clay and silt)
PB-24	Top of semi-confining unit encountered at 140 ft bg…End of drilling at 150.5 ft

Table 1.1: Excerpts from drilling logs indicating the presence or absence of facies which could be the Gardiners Clay

Table 1.1 shows those sections of some drilling logs that indicate Gardiners Clay, or not, at five boring locations. The methods used in these descriptions are not consistent, and the drilling logs are variable in reporting the observations. The information in the excerpts is important for representing the stratigraphy of the model domain with reasonable accuracy; however, the variability in the linguistic descriptions hamper its use. Although all descriptions are in reference to the same geologic unit, the difference in the descriptions add to input uncertainty.

Reasonably accurate interpretation of the hydrogeology of the study area is important for defining a conceptually sound model domain (Reilly and Harbaugh 2004; Buxton and Reilly 1985). The quality of the modeler's conceptual model will depend on the expanse and quality of sampling data relative to the spatio-temporal scale of the real world hydrogeologic system. The data's inability to represent the system that, in turn, cause imperfect understanding of the system in question.

1.4.1.3. Parameter Uncertainty

A major difficulty in groundwater modeling is that sub-surface features of the groundwater system – the structure of the aquifer and aquifer characteristics – are hidden. Geologic layers may be discontinuous in their arrangement because of the re-working, such as by erosion or glacial events. For example, the Matawan Group-Magothy Formation sedimentary deposits occupy most of Long Island but the surface of this layer is severely eroded as a result of advancing and retreating glaciers and glaciofluvial channeling (Sirkin 1982). Hydraulic connections between aquifer units may be disturbed because of presence of localized or widespread aquitards. For example, geologic borings indicate the presence of discontinuous zones of solid clay of variable thickness that form localized clay lenses in the Matawan Group-Magothy Formation on Long Island (Smolensky and Feldman 1992). The rate of recharge may be different for deeper aquifers than for shallower aquifers. For example, the approximate travel time of water recharging at the regional groundwater divide ranges from 25 to 100 years for the Magothy aquifer and from 400 to 3,000 years for the Lloyd aquifer. Hydrologic characteristics within an aquifer may vary through the model domain in important and meaningful ways. For example, the values of the hydraulic conductivity in the Upper Glacial aquifer depend directly on grain size and indirectly on depth. The deposits in the upper sections are generally coarse and readily yield water, while the deeper sections have better sorted, layered sands with low permeability. Generally, the hydraulic conductivity in the deeper sections Upper Glacial aquifer is about one-third of that of the upper sections of the Upper Glacial aquifer (McClymonds and Franke 1972). Changes in characteristics may be heterogeneous and variable, and discontinuous. For example, the presence of clay layers in the Matawan Group-Magothy Formation causes a high degree of anisotropy locally. Therefore, the groundwater systems are not only hidden, heterogeneous and complex, but they can also be non-linear in their response to stress; for example, as in the unconfined aquifers where the head could be a non-linear function of stress (Jansen 2003; Beven 2001).

Parameter uncertainty is associated with the data and the methods used for model parameters (Walker et al. 2003). Typically, hydrogeologic regime of the study areas are characterized using data such as geologic borings, water quality indicators, streamflow data, and groundwater head measurements. Some of these data, such as the geologic boring logs, are estimated for a point in the study area and then extrapolated over larger, heterogeneous areas

(Wagener and Gupta 2005). Also, it is difficult to measure all the observational data at time scales that match the level of detail at which the model outputs are simulated (Lane and Richards 2001). Therefore, groundwater models are constrained with regard to the number of observations that are used to develop the model structure and conceptualization. Also, non-linear processes, heterogeneous model domains, measurement errors, and the spatio-temporal limitations on empirical data collection make it challenging to develop models that incorporate the unique characteristics of a particular groundwater regime.

Therefore, the complexity of reality is typically represented by parameters, that is, aggregate approximations lumped together in space and/or time (Waganer and Gupta 2005). The parameters are expected to represent the average value/state/behavior of the large-scale, heterogeneous hydraulic properties of the system within a structural grid cell (Refsgaard et al. 2012). Parameters values are derived from many sources: historical site specific data, non-site specific or surrogate data, or from parameter calibration programs. For example, hydraulic conductivity (K) of an aquifer is not directly measured in a Darcy experiment but it is inversely derived from the observed data as follows (Equation 1.3):

$$K = -\frac{Q}{A}\left(\frac{dl}{dh}\right) \tag{1.3}$$

Where,

Q = the discharge ($L^3\ T^{-1}$),

A= the cross-sectional area of the aquifer (L^2),

dh = the change in head between two closely placed points (L^1), and

dl = the distance between these points (L^1)

All variables on the right hand side of the equation are directly measurable in a Darcy experiment. This type of solution is described as an inverse problem, where the values of model components are derived from measured quantities. Inverse problems are different from forward problems where the values of the model components are known and these are used to determine the value of the system response quantity such as groundwater heads (Anderson and Woessner 1992, p. 226; Wang and Anderson 1992, p. 45).

Field experiments conducted in the landfill vicinity indicated that the horizontal hydraulic conductivity (K_h) values for the Upper Glacial aquifer varied from 17 feet/day to 1,437 feet/day,

depending on the sections of the aquifer sampled and the sampling technique used (Dvirka and Bartilucci 1994a) (Table 1.2). Wexler and Maus (1988) used a parametric value of 267 feet/day to represent the K_h of the Upper Glacial aquifer in the landfill model; this value was based on a previously published K_h estimate for the Upper Glacial aquifer (McClymonds and Franke 1972).

UGA Zone	Monitoring well	Screened Interval Depth (ft below grade)		Hazen method geometric mean K_h (ft/d)	Bouwer-Rice Rising Head Test K_h (ft/d)	Bouwer-Rice Falling Head Test K_h (ft/d)
		from	to			
Shallow	MW-1S	38	58	116	116	--
	MW-6S	38	58	119	417	--
	MW-3S	39	59	128	607	--
	MW-4S	39	59	95	94	--
	MW-10S	40	60	112	1074	1437
	MW-5S	41	61	95	411	--
	MW-9S	41	61	119	369	--
Intermediate	MW-5I	80	90	208	67	69
	MW-8I	80	90	158	187	146
	MW-10I	80	90	148	58	155
Deep	MW-2D	129	139	56	79	29
	MW-10D	140	150	58	19	19
	MW-6D	143	153	51	100	31
	MW-4D	150	160	48	363	131
	MW-8D	158	168	30	17	20

Table 1.2: Horizontal hydraulic conductivity (K_h) values for the Upper Glacial aquifer (Dvirka and Bartilucci 1994a)

Table 1.2 suggests that there is noticeable variability within a single aquifer not only with respect to the locations but also with respect to measurement methods used. A parameter value that aggregates variabilities may not produce good results if the local scale heterogeneity in K_h is needed for an accurate simulation (Binley et al. 1989). Judgments regarding the sources of parameter values are useful in selecting more representative parameter values, but they do not compensate for the aggregation of the heterogeneities of these parameterized variables. In addition, site specific data may not be spatio-temporally comprehensive enough to cover the entire range of the parameterized variable. The scale of the surrogate data may not be

comparable and the association between the model domain and the site at which the surrogate data were derived is difficult to establish. Calibration procedures choose parameter values that reduce result discrepancies between the model and the reality, but one should ensure that the calibrated parameter values are reasonable and consistent with what is understood about the hydrogeologic reality (Jansen 2003). The scale of real world observations is usually finer than the scale of model's grid given that the grid cells are typically order of magnitude larger in dimensions (a few to several hundred square feet) compared to the dimensions of an observation well (a few inches in diameter). This discrepancy also affects the representativeness of the parameterized value (Jansen 2003).

1.4.2. Levels of Uncertainty

Model uncertainties are unequal in magnitude and occupy different positions on an uncertainty spectrum that shows levels of uncertainty (Figure 1.8).

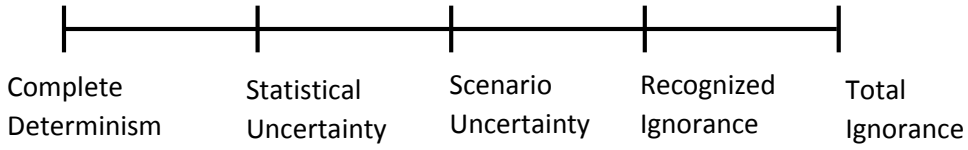


Figure 1.8: Uncertainty spectrum showing levels of uncertainty

Determinism is the ideal, but impossible, situation where the modeler possesses the complete and precise knowledge of groundwater system. Statistical uncertainty arises due to inaccurate and imprecise measurements. It is any uncertainty that can be adequately characterized in statistical terms, and is applicable to any location in the model, including model structural uncertainty. Scenario uncertainty indicates the failure to incorporate the probability of occurrence of particular scenario from multiple possibilities. It arises because of lack of understanding of the mechanisms that cause the manifestation of that scenario. Recognized ignorance is uncertainty about the fundamental functioning and mechanisms of the groundwater system. It can be reduced by additional empirical research and data collection provided that these additional data address the indeterminate elements of the groundwater system. Finally, total ignorance is the total opposite of determinism; it implies the state of "unknown unknowns": we do not know that we do not know (Walker et al. 2003). It is to be noted that fixing the level of an uncertainty is a theoretical construct because it would require exact knowledge of how much

uncertainty exist. The different levels of uncertainty are qualitative and therefore are subject to change depending on their interpretation by different modelers.

1.4.3. Nature of Model Uncertainty

The third type of uncertainty classification is based on the nature or the thematic cause of model uncertainty. This classification focuses on the thematic causes of the rise of different model uncertainties. There are two classes: "epistemic uncertainty" and "aleatory uncertainty" (Rozell and Reaven 2011; Roy and Oberkampf 2011; Karnaki et al. 2009; Ferson et al. 2008; Brugnach et al. 2007; Oberkampf et al. 2002).

Epistemic uncertainty arises because of the lack of knowledge about the characteristics and behavior of the groundwater system that is being modeled. Epistemic uncertainty is also known as "incertitude", "ignorance", "subjective uncertainty", "non-specificity", "reducible uncertainty", "secondary uncertainty", or, "cognitive uncertainty" (Oberkampf et al. 2002). Some causes for epistemic uncertainty are: measurement error, small or limited sample sizes, detection limits ("non-detects"), data censoring, missing values, use of surrogate data, ignoring the details of physical mechanisms, imperfections in scientific understanding, rounding error, intermittent measurement of a periodic process, subjective judgments, and ambiguities. Epistemic uncertainty can be reduced by additional data collection and empirical research, improved numerical approximations, and with expert consultation.

Aleatory uncertainty arises because the effect of chance. It is a function of natural stochasticity of the system. Aleatory uncertainty is also known either "randomness", "variability", "stochastic uncertainty", "objective uncertainty", "dissonance", "inherent uncertainty", "primary uncertainty", or "irreducible uncertainty" (Oberkampf et al. 2002). Some causes of aleatory uncertainty are: inherent variability of the system, environmental or structural variation across space and time, or heterogeneity among components, external input data and functions, parameters, and model structures. Unlike epistemic uncertainty, aleatory uncertainty may not be reduced by additional data (Walker et al. 2003). It may be reduced if the real world system itself is changed to a state where the inherent fluctuations no longer exist or that they can be explained (Karnaki et al. 2009).

All uncertainties in groundwater models can be classified either as epistemic uncertainty or aleatory uncertainty. For example, conceptual uncertainty is an epistemic uncertainty, because

it can be reduced with greater understanding of the groundwater system (Singh et al. 2010; Rojas et al. 2008). Parameter uncertainty is also on epistemic uncertainty because parameters can be more accurately characterized with adequate data (Karnaki et al. 2009). On the other hand, the fluctuations in groundwater levels is an inherent variability in the observation data and therefore it is classified as an aleatory uncertainty. This variability can exist among individuals within and across populations of the observed data, and also within an individual observation over time. Better observational tools and techniques can reduce the measurement error and make these measurements more accurate. The degree of fluctuation can be better in these measurements can be better characterized by collecting more data such as historical records of groundwater heads, or location specific, real-time information about recharge, runoff, and evapotranspiration. However, sufficiently long term records may not be available for all these variables at all well locations. As a result, it become difficult to accurately characterize the fluctuations in the groundwater levels. Hence, the variability in the groundwater heads can be classified as aleatory uncertainty.

Examination of epistemic uncertainty is necessary to estimate the value of conducting additional experimental or empirical research and data collection (Morgan and Henrion 2006, p. 63). The purpose of additional, periodic, and more dispersed data collection is to reduce epistemic uncertainty. The cost associated with additional data collection should be weighed against the benefits that accrue by reducing model uncertainty. Identification of reducible uncertainties and non-reducible uncertainties optimizes additional data collection exercises. In some cases, model uncertainties could remain if they are acceptable for the purpose of the model or if it is infeasible to rectify them at a reasonable expense (Oberkampf et al. 2002). Examination of aleatory uncertainty may be necessary to estimate the relative worth of separating these two types of uncertainties while assessing the performance of the model (Morgan and Henrion 2006, p. 63).

1.5. Multiple Models

1.5.1. Developing Multiple Models

Traditionally, groundwater models have been constructed on the basis of a single geological model structure that is assumed to be the best possible representation of the groundwater system. It has been argued that this approach fails to sample the complete space of plausible models adequately and therefore undervalues uncertainty (Neuman 2003; Neuman and Wierenga 2003). Typically, a large number of inputs are included in a complex, physically distributed groundwater model and each of these inputs may have an uncertainty attributed to them. This hyper-dimensional uncertainty cannot be propagated through a unique, deterministic model solution with fixed model settings (Konikow 1996). This leads to the problem of model "non-uniqueness": given model uncertainties, no particular single model can be deemed as the unique modeling solution to the problem at hand.

As a remedy, multiple models can be developed and tested if the model conceptualization is ambiguous, if the hydrogeologic framework of the study area is uncertain, and if the observational data used in model development are limited (Voss 2011; Rojas et al. 2008; Refsgaard et al. 2006). Theoretically, developing multiple models to explain the same real-world system is analogous with Thomas Chamberlin's argument against forming a positive bias for a particular ruling theory and instead developing "multiple working hypotheses" to explain the real world phenomenon (Chamberlin 1965). Chamberlain argued that the investigation would "lack completeness" if only a single working hypothesis is pursued while neglecting other, equally plausible hypotheses, and one should therefore do the otherwise.

Similarly, if the model conceptualization is ambiguous, if the hydrogeologic framework of the study area is uncertain, and/or if the observational data used in model development are limited, then multiple models should be developed instead of a single model (Voss 2011; Rojas et al. 2008; Refsgaard et al. 2006). Modelers can thus test alternate model conceptualizations, and not be restricted to one particular model. Multiple models are generated based on varying combinations of the inputs, parameters, and conceptualizations, as well as using various computational schemes (finite-difference or finite-element) or dimensionality (1-D, 2-D, or 3-D), to incorporate model uncertainties. The use of multiple models increases the reliability of simulated outputs estimates, because there is explicit incorporation of model uncertainties (Ye et

al. 2010). The use of multiple unique combinations of model conceptualizations and parameters may allow the mapping of the uncertain input space to the uncertain output space (Roy and Oberkampf 2011). Thus, inferences derived from multiple models are considered to be more realistic and robust given the goodness-of-fit of multiple model representations are tested instead of being faithful to a singular model (Poeter and Anderson 2005).

The process of generation of multiple models can be readily automated if the model variations are limited to varying quantities of input parameters and variables, with a fixed model framework. Stochastic versions of MODFLOW (Environmental Solutions Inc.) can randomize the generation of parameter sets using Monte Carlo simulations. Where the uncertainty in model includes geological uncertainty, model versions with each version a unique combination of model conceptualization, parameters, and inputs can be created (Refsgaard et al. 2012). For example, TPROGs (Transition Probability Geostatistical Software) (Carle 1999) uses geostatistical approaches such as transition probability and Markov chains for stochastic generation of alternate geologic structures.

Most commonly, multiple models are generated manually. The modeler generates a wide array of potential combinations of the model features derived from sources including empirical measurements, data collection, other simulations, and expert opinions (Roy and Oberkampf 2011). Obviously, such flexibility should be bound by the basic principles of the sound modeling practices, practical constraints, and by the purpose of the modeling exercise (Reilly and Harbaugh 2004). It is understood that this approach generates a pool of models that may not be exhaustive and may be biased and/or under-dispersive compared to the true model space. Perhaps, the generation of the structural framework could be randomized by treating the framework as an uncertain parameter, and then selecting and subsequently simulating a range of randomly selected frameworks from a pre-defined distribution of frameworks. A Bayesian-type analysis could be used in such case where the distribution of the uncertain framework parameter can be updated based on the models' performance. In the present study, a simpler, manual and combinatorial approach was preferred to generate the multitude of model variants; this approach gave better control over the validation assessment and also made the computation and the following analysis time-efficient.

1.5.2. Testing Models (Model Validation)

When multiple models are generated, the modeler needs to identify the models that best match the real world system in terms of the system's observable behavior. Models can be evaluated by a process, model "validation". Validation is defined as determining the degree to which a model is an accurate representation of the real world system, from the perspective of its intended uses (Ferson et al. 2008; Oberkampf and Barone 2006; Law 2005; Law and Kelton 2000, p. 264; Refsgaard and Knudsen 1996). Validation is also known as "verification" (Lane and Richards 2001), "accreditation" (Balci 1998), "evaluation" (Oreskes et al. 1994), or "assessment" (Gass 1983).

Model validation is classified into four constituent components: conceptual validation, model code verification, data validation, and operational validation (Sargent 2009). Conceptual validation addresses subjective choices of parameters, inputs, and processes in the model. Model code verification is the evaluation of the numerical formalization of the conceptual model. Data validation or quality control tests the check for the accuracy, unbiasedness, adequacy, and correctness of the data (Sargent 2009; Schellenberger 1974). Operational validation determines whether the model's output behavior is sufficiently accurate for the model's intended purpose. Operational validation is further classified as (i) replicative validation: a demonstration of the model's ability to replicate existing observational data and, (ii) predictive validation: a demonstration of the model's ability to accurately forecast trends and values of the real world system (Gass 1983). Validation can either be qualitative, using subjective assessment tests, or quantitative, using numerical techniques. A number of validation schemes have been proposed (Balci 1998; Forrester and Senge 1980). No single validation can guarantee a model's representativeness. Results from various performance, uncertainty, and subjective assessment tests can be synthesized to determine a model's overall adequacy (Barlas 1996; Forrester and Senge 1980).

Replicative validation is commonly used. It is a quantitative assessment of a model's representativeness (its goodness-of-fit), conducted by comparing predictions or simulation results of the model with corresponding empirical data or experimental measurements (Roy and Oberkampf 2011; Ferson et al. 2008; Romero 2007; Truncano et al. 2006; ASTM 2002, Standard D5490; Oberkampf and Barone 2006; Hills 2006; Oberkampf et al. 2004; Oberkampf and Truncano 2002).

The validity of a groundwater simulation model can be assessed by comparing the groundwater head observations made at wells in the model domain with the groundwater head values simulated by the model. Suppose a single head observation is available (n=1) and a single corresponding head is simulated by the model. Here, the replicative validity of the model is assessed by measuring the difference (*d*) between the observed datum and the simulated datum (Figure 1.9). Smaller *d* values suggest better agreement between the model and the real world system and thereby, better replicative validity.

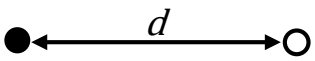


Figure 1.9: Difference (*d*) between the observed datum (solid circle) and the simulated datum (hollow circle)

When multiple models are simulated, uncertainty leads to different models simulating different outputs. The degree of difference between the observed datum and the simulated datum generated by each model can then be compared to assess each model's replicative validity; in this example $d_{M2} < d_{M1} < d_{M3}$ (Figure 1.10).

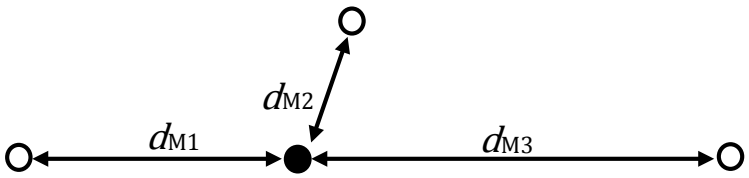


Figure 1.10: Difference (*d*) between the observed datum (solid circle) and the simulated datums from models M1, M2, and M3.

Given a set of multiple models, a "multi-model analysis" (MMA) can be carried out to assess the models' replicative validity. Many such procedures have been developed. Here, four MMA procedures are described for illustrative purposes. These procedures are: model selection (section 1.5.2.1), model averaging (section 1.5.2.2), multi-objective optimization (MOO) (section 1.5.2.3), and Generalized Likelihood Uncertainty Estimation (GLUE) (section 1.5.2.4).

1.5.2.1. Model Selection

Model selection involves testing the performance of multiple models to increase the chances of bracketing the true model (Rojas et al. 2008). For example, the Akaike Information Criterion (AIC) (Akaike 1973) is a relative optimization method that uses the maximum

likelihood approach to rank model performances using Akaike weights as a metric. The AIC-based model selection approach penalizes model complexity, defined as the number of parameters included in the model. This approach selects the most efficient model: the model that achieves the best fit with the observed data using the fewest parameters (Poeter and Anderson 2005). AIC is calculated as follows (Equation 1.4):

$$AIC = n \ln(\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}) - n \ln(2\pi) - n + 2p \quad (1.4)$$

Where,

n = number of observations,

p = number of estimated parameters for the model +1, and

ε = residuals (observed minus simulated values).

The difference (Δ_i) between the AIC of a particular model i (AIC_i) and the model with the least AIC value (AIC_{min}), Δ_i, is obtained (Equation 1.5). Larger Δ_i indicates that it is less likely that the model will be the most efficient one.

$$\Delta_i = AIC_i - AIC_{min} \quad (1.5)$$

Akaike weights, w_i, are derived for each model from the Δ (Equation 1.6). These weights indicate the relative probability of a model with respect to the family of n models.

$$w_i = exp^{(-0.5\,\Delta_i)} / \sum_{i=1}^{n} exp^{(-0.5\,\Delta_i)} \quad (1.6)$$

A model receiving a higher Akaike weight is deemed the model most likely to be representative of the real world system. Other approaches based on information criteria, such as the Bayesian Information Criterion (BIC), or the Kashyap Information Criterion (KIC) are also used for model selection (Poeter and Anderson 2005).

The disadvantage of this approach is that the model weight obtained with respect to one observational data set cannot be compared with the weight obtained with respect to a different observational data set (Anderson and Burnham 2002). This limits the ability of the model selection process to evaluate the same model with respect to different observational data sets representing a range of system behavior.

1.5.2.2. Multi-model Averaging

In multi-model averaging, the representativeness of different models is assessed on the basis of a performance metric; then the simulated output from these models is weight-averaged using the performance metric as the weight (Engelhardt et al. 2012, Ye et al. 2010, Poeter and Anderson 2005). For example, the multi-model average of a simulated output can be computed using the Akaike weights as the performance metric as follows (Equation 1.7):

$$S_{avg} = \sum_{i=1}^{n} w_i S_i \quad (1.7)$$

Where,

S_{avg} = multi-model average value for parameter or prediction estimate,

w_i = Akaike weight for the model i (i=1,…, n) from Equation 1.6 above, and

S_i = simulated output for model i (i=1,…, n).

Likewise, Bayesian Model Averaging (BMA) employs probabilistic techniques to derive consensus outputs from a set of alternative models (Hoeting et al. 1999; Draper 1995). BMA weights the simulated outputs of competing models by their corresponding posterior model probability, representing each model's relative success in reproducing system behavior in the training period, the length of the period used to estimate the BMA weights. Similarly, the Maximum Likelihood BMA (MLBMA) is an approximation of BMA. It relies on the maximum likelihood parameter estimation and assesses the joint predictive distribution of several competing models (Neuman 2003). MLBMA does not require exhaustive Monte Carlo simulations and obviates the need of prior information about model parameters, which is often difficult to obtain (Ye et al. 2005). Posterior model probabilities are subsequently approximated using the Kashyap Information Criterion (KIC) (Kashyap 1982).

Model averaging methods are useful to generate a composite of the results derived from multiple models because these methods aggregate multiple models using weighted averages and generate a singular estimate. However, all of the models need to be composed of identical parameter structure to enable the weight averaging. The disadvantage of this large parameter-perturbed ensemble is that it does not sample structural uncertainty in model configurations because of use of a singular framework. Models with varying hydrogeologic framework may have variable numbers of parameters. The aggregation may incorporate a wide array of models

that may be significantly dissimilar from one another making it difficult to reconcile different model configurations into one single model structure. A singular model-simulated response can be generated using the ensemble averaging method, but such a disjoint ensemble may not be reconciled into a singular and meaningful model configuration to be used for forward problem solving. Thus, there are practical difficulties in aggregating disjoint models in an ensemble.

1.5.2.3. Multi-model Optimization (MOO)

Judging model performance by one type of observed data may not be appropriate when multiple sets of observed data are available (Yapo et al. 1998). For example, a better model would perform well with respect to both groundwater head and streamflow discharge, compared to models that only do well with respect to one data set. Minimization (or maximization) of individual performance measures may result in non-unique model solutions.

Multi-objective optimization (MOO) is based on the principle of model equivalence: more than one model can be representative of the real world system as determined by multiple and incommensurate measures (Gupta et al. 1998). MOO techniques simultaneously minimize (or maximize) more than one performance measures with respect to parameter set θ (Demarty et al. 2004) (Equation 1.8):

$$\text{Min } \{F_1(\theta), \ldots, F_n(\theta)\} \tag{1.8}$$

where, $F_n(\theta)$ = n different performance measures

MOO bisects the model set into a "behavioral" set and a "non-behavioral" set. A model in the behavioral set should be better than other models in that set with respect to all objectives except for one, at least. It is not possible to identify a model from the behavioral set that outperforms all the other models in all performance measures; there is always a trade-off among the performance of a particular model with respect to different performance measures (Gupta et al. 1998). For example, suppose two performance measures are to be minimized using two parameters (Figure 1.10-a). The solution at point A is the solution such that it minimizes objective 1, while the solution point B is the solution that minimizes objective 2. The curve joining point A and B is the Pareto front. The solutions that lie on the behavioral front are the Patero set solutions (Yapo et al. 1998) (Figure 1.11-b).

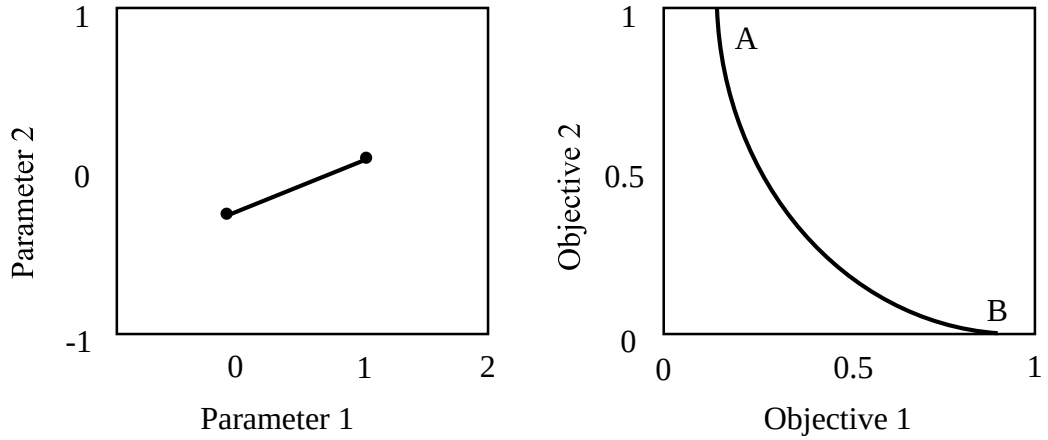


Figure 1.11: (a) parameter space, (b) objective space (Yapo et al. 1998)

A number of automated MOO procedures can be used, such as the Shuffled Complex Evolution (SCE-UA) (Duan et al. 1992), and a multiple criteria global optimization algorithm (MOCOM-UA) (Vrugt et al. 2003).

MOO methods are exclusively applied for parameter optimization under a fixed model conceptualization. For example, Efstratiadis and Koutsoyiannis (2010) reviewed 36 MOO applications in hydrological modeling; only the number of parameter sets and the objectives were altered, while other features of the model framework remained fixed. Therefore, although the MOO is based on a rigorous model selection process, achieving a closer model fit via parameter adjustment may overcompensate for a potentially erroneous conceptualization. Therefore, the assumption of an error-free model conceptualization is not justifiable.

1.5.2.4. Generalized Likelihood Uncertainty Estimation (GLUE)

The theory of model equifinality postulates that it is impossible to identify a unique true model given limited observed data, and so that more than one model could be deemed to be representative of the real world system (Romanowicz and Beven 2006; Beven 2006, 1989; Beven and Binley 1992). The model equifinality thesis forms the basis of the Generalized Likelihood Uncertainty Estimation (GLUE) procedure (Beven and Binley 1992). The GLUE procedure bisects the multi-model set into behavioral and non-behavioral models, similar to MOO, on the basis of a rejection threshold that uses a likelihood measure of the modeler's choice. The GLUE likelihood is an indication of the degree of membership in the behavioral model set; models receiving higher likelihood values have higher chances of being representative

(Freer et al. 1996, Beven and Binley 1992). The GLUE likelihood is "generalized" because the choice of the performance measure to evaluate the model likelihood is subjective (Jansen 2003; Beven and Binley 1992). A number of different performance measures are used as GLUE likelihood measures (Table 1.3).

Performance measure	Formula
Root mean squared error (RMSE)*	$\sqrt{\frac{\sum_{i=1}^{N}(S_i - O_i)^2}{N}}$
Exponential function *	$\exp\left(\frac{-M\sum_{i=1}^{N}(O_i - S_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2}\right)$
Inverse error variance *	$\left(\frac{\sum_{i=1}^{N}(O_i - S_i)^2}{N}\right)$^-Z
Model efficiency (ME)	$\left(1 - \frac{\sum_{i=1}^{N}(O_i - S_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2}\right)$^M
Index of Agreement (IoA)**	$1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(\lvert S_i - \bar{O}_i\rvert + \lvert O_i - \bar{O}_i\rvert)^2}$

Table 1.3: Performance measures used as GLUE likelihood measures (S_i = simulated value, O_i = observed value, $\bar{O}$ = mean observed value, $\bar{\mathbf{S}}$ = mean simulated value, N = number of measurements, M = scaling factor) (*Stedinger et al., 2008, Jansen 2003, Beven and Binley 1992, ** Legates and McCabe 1999; Janssen and Heuberger 1995, Loague and Green 1991, Willmott 1981.)

Note: ME is equivalent to Nash-Sutcliff efficiency (NSeff) or Coefficient of Determination (CoD) when M=1. All models are equiprobable when Z=0 and the likelihood response surface converges to a single peak or a single best model when $Z \rightarrow \infty$.

The GLUE likelihood implicitly incorporates different sources of uncertainty (conceptual, parameter, and input) (Beven and Freer 2001). It is used to develop multi-model weighted averages of the parameter values (Singh et al. 2010; Romanowicz and Beven 2006; Romanowicz and McDonald 2005; Jansen 2003; Beven and Binley 1992). In addition, the GLUE likelihood can incorporate additional information via Bayesian updating to form posterior likelihood for a given model (Romanowicz and McDonald 2005). The disadvantage of GLUE

procedure is that the GLUE likelihood measures and the rejection criterion that is used to classify a model as behavioral or non-behavioral are subjective and not standardized. It is suggested that a rigorous rejection threshold should be used to minimize this subjectivity (Todini and Mantovan 2007; Mantovan and Todini 2006).

1.5.3. Limitations on the MMAs

The above mentioned MMAs are of limited use for two reasons:

a. There is epistemic and aleatory uncertainty associated with the observed data; and

b. The simulated data are a result of a simplified representation of the real world system

1.5.3.1. Uncertainty Associated with the Observed Data

Observed data, such as groundwater heads, vary among different wells, as well as vary over time at an individual well. The fluctuation of observed data can be bracketed by an interval range; the exact value of these data is not fixed given the variation in the observed data over time. These fluctuations in observed data are classified as aleatory uncertainty, one associated with the groundwater flow system. For example, Figure 1.12 shows aleatory uncertainty measured at well S3529 located near the Brookhaven landfill for the period 1975-2010. The water table fluctuated by nearly 8 feet, from a minimum of 22.32 feet (November 2002) to a maximum of 30.20 feet (June 1998). In 1998, the head rose by about 5 feet (from 25 feet in January to more than 30 feet in June) and again fell by nearly 4 feet (from 30 feet in June to around 26.5 feet in December). At times, the water table fluctuations were relatively small, such as from 1986 to June 1989.

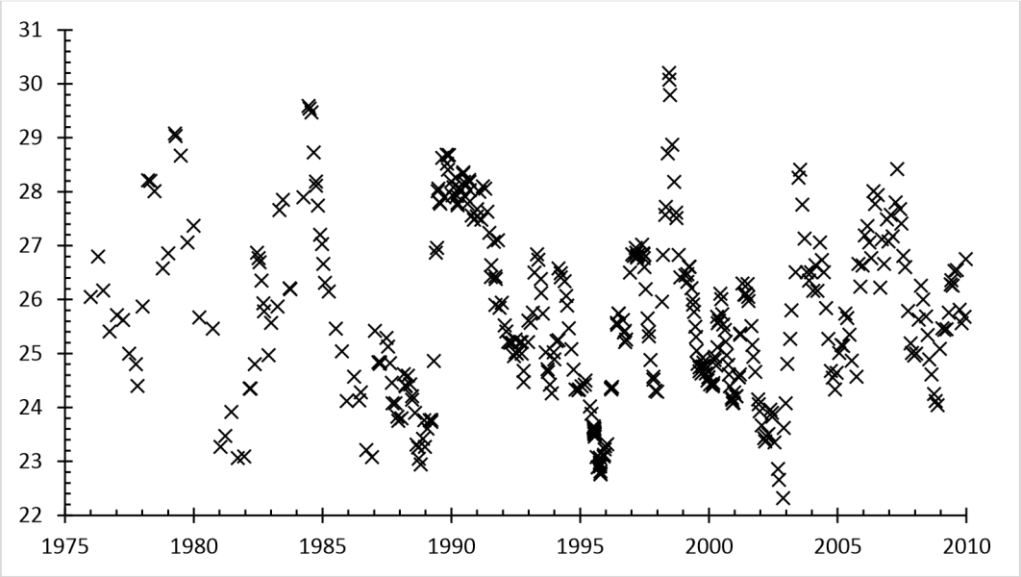


Figure 1.12: Water Table Heights (in feet msl), Well S3529 (1975-2010)

Generally, groundwater heads fluctuate spatio-temporally due to differences in precipitation trends, the propagation of long-term changes in recharge from upgradient to downgradient, as well as due to short-term responses to the precipitation events preceding the measurement (Romanowicz and MacDonald 2005; Peterson 1987; Steenhuis et al. 1985; Franke and McClymonds 1972). Fluctuations can also be caused from an artificially introduced stress such as pumping, changes in the barometric pressure, and for a confined coastal aquifer, changes in the hydrostatic pressure induced by tidal fluctuations (Brassington 1988, p. 80-81).

On the other hand, each individual datum has an epistemic uncertainty associated with that observation. Epistemic uncertainty is associated with measurement error from faulty devices, mistakes by the observer while registering the measurement, technical error where the accuracy of the instrument is coarser than the resolution of the observation (non-detects), and from rounding errors (Romanowicz and MacDonald 2005). Restricted access to the observation point due to flush-mounted wells, brush, or protective casings, or installing pumps in the borehole can make clear head observations difficult (Brassington 1988, p. 75). Well casings can become damaged or otherwise de-leveled, changing the measuring point elevation. Tight caps, especially when wells are screened in low permeability strata, can require a long time for the casing water level to equilibrate to atmosphere conditions.

Because of epistemic and aleatory uncertainty associated with observed data, a static, scalar value for observed data is a "snapshot" and an uncertain representation of the groundwater system. Sometimes, the mean observation value is used for model validation; however, the mean represents only the average behavior and not the range of behavior displayed by the system. Therefore, the mean or other measures of central tendency may not be sufficiently informative for the intended use of the model (Ferson et al. 2008). Automated parameter estimation methods allow models to be deemed valid, based on static comparisons between the observed data and the simulated values. Such models may be overfitted ("tuned") to one particular dataset, one that may represent just one state of the system. Such models tend to perform poorly when tested against data set from a different time period (Konikow 1996).

1.5.3.2. Simplified Representation of the Real World System

The system response quantity simulated by one particular model is produced by that model that, in turn, is the product of a set of model inputs. These deterministic values are derived from a model that is a product of assumptions, simplifications, and lumped approximations. Therefore, the correspondence between the real world conditions and the model conditions is unlikely to be exact. This correspondence may be poor given the discrepancies in scale of input information, uncertainty in input data, missing data, lack of location specific information, and discretization errors. A model can be considered a robust representation of the real world system if model components are all completely and deterministically specified (Oberkampf and Roy 2010, p. 98). However, the determinism in the model simulated output is questionable. Therefore, a close match between the observed and the simulated values may give a false impression of certainty. In addition, the model is spatio-temporally discretized and this discretization is necessarily coarser than the continuous space and time of the real world system. So, it has been said the observed and simulated values are incommensurate because they have different spatio-temporal scales (Beven 2012).

Models are usually simulated to portray a particular state of the real world system. Groundwater simulation models generate deterministic output quantities at the locations of the observation wells. Therefore, no probabilistic value accounting for uncertainty is attributed to the model simulated output; each output is considered as a fixed, deterministic quantity and the differences between the observed values and the simulated values are exact. But such precision

in the results do not reflect the underlying model uncertainty. Models that fit a particular state may not have much utility because model predictions may not be needed for only that particular state. Therefore, an emphasis on deterministic replicative validity is not fruitful. So, assessments of the predictive accuracy of traditionally validated groundwater models have found that they are not often accurate (Konikow 1996; Anderson and Woessner 1992, p.69).

A model consistency, with respect to the system behavior, has been preferred over the identification of a single best model via optimization methods (Wagener and Gupta 2005). A model that suits multiple system states has more value because it provides addition proof that the model holds true over a range of different assumptions and future scenarios, and not over a particular set of observed data (Neuman 2003). As the number of observations increase, to generate one unique model to suit every single state of the system may become infeasible and unwarranted. For example, with respect to the observed data at well S3529, two separate models can be simulated, one to simulate conditions for the highest groundwater head recorded (30.20 feet) and the other to simulate the lowest groundwater head recorded (22.32 feet); however, a model that simulates a value of, say, 26.61 feet may be considered a reasonable representation of measured conditions for head values 26.71 feet and 26.50 feet. The adequacy of the model's goodness-of-fit with the reality is obviously subjective, but given the uncertainty involved in modeling, an attempt to achieve a point-by-point match could undermine the goal of groundwater modeling, as described by Voss (2011), that "*is to learn and to gain additional insight about the system under study rather than attempting to achieve detailed data fitting of highly parameterized models by merely pushing buttons and adjusting knobs on a computer program*".

1.5.3.3. Summary

Acknowledging the uncertainty in model development is an important step to ensure a model is not presented as a complete understanding of the system, but as a scientific expression of our ignorance (Doherty 2011). Testing model uncertainties using multi-model analysis, rather than assessing the appropriateness of a singular model, is preferable because the model user has assurances uncertainty has been explicitly incorporated using multiple model conceptualizations. Two factors limit the validating abilities of the traditional multi-model analyses. First, the observed data used in these analyses could be uncertain given the inaccuracies in the

observational tools and techniques. The observed data, such as groundwater heads, may inadequately represent the spatio-temporal heterogeneity of the real world system because these data are not constant but fluctuate through space and time. Second, models are not exact replicas of the real world system, but are simplified versions formed from assumptions, extrapolations, and discretizations. Simplifications are made with respect to model settings. For example, aquifer properties, such as the hydraulic conductivity of an aquifer at various locations, are lumped into single parameter values. Historical data and surrogate data are used as inputs. The continuous terrain and geology are discretized into finite-grid model domain, while time is aggregated into coarse steps. Therefore, although the model simulated values are deterministic and have a one-on-one correspondence with the observed data, it seems unlikely that the model accurately represents the exact state of the system when the observations were made. For these reasons, the traditional approaches for validation assessment based on deterministic comparisons should be revised. Validation tools and methodologies that account for heterogeneities, uncertainty, and the scale of data should be developed and their validity be assessed.

Chapter 2

Objective

The objective of this study is to conduct a multi-model validation assessment through an approach called the "area metric" (Ferson et al. 2008). In the area metric approach, the observed and the simulated data are expressed as the empirical cumulative distribution functions (ECDFs). The ECDF of the observed data is a graphical summary of the various states of the groundwater system as reflected by the system's response quantity – in this case, groundwater heads. Likewise, the ECDF of the simulated data is a compilation of a set of deterministic values generated by simulating a given model a specified number of times to represent different system states. The area metric indicates the difference between these two quantities, calculated as the area between the ECDF of the observed values and the ECDF of the simulated values.

The area metric-based multi-model validation assessment will be applied to multiple variants of a base model. This base model simulates the groundwater conditions in the vicinity of a municipal solid waste landfill site in the Town of Brookhaven in Suffolk County, New York (the study area). My hypothesis is that this approach facilitates selection, from the model space, of a set of models that show a better agreement with the real world groundwater conditions observed in the study area.

The groundwater system in the study area is a representative example of a complex, physically distributed, and spatio-temporally heterogeneous system. The configuration and the characteristics of this system are uncertain. These uncertainties present a challenge in developing representative simulation models of the system. As a result, the hydrogeologic framework and the characteristics of the study area was depicted by multiple candidate models representing the same real world system. These models were developed and simulated, and their replicative validity was assessed using a two-step process involving the proposed approach.

The effective definition of validation followed here is replicative validation: "validation is an assessment of model accuracy by way of comparison of simulation results with experimental measurements" (from Roy and Oberkampf 2011). An assessment of the replicative validity between the observed data and corresponding model simulated outputs was made through their respective ECDFs. The observed data ECDF was developed using the groundwater head observations measured at numerous wells distributed across the study area. The simulated data ECDFs were developed using the groundwater head data simulated by the different model variants at these well locations.

Numerous model features are included in these complex, physically distributed models and each feature has its own type, level, and nature of uncertainty. Here, the model features were classified into two groups; the variable features, features whose uncertainty was acknowledged and incorporated into the model, and fixed features, whose uncertainty was acknowledged but not incorporated into the model. There were eight variable features. The remainder of the model features were kept fixed. This select set of models is not exhaustive; it is likely that several different model conceptualizations can be assessed through the multi-model analysis if uncertainty in additional model features is incorporated.

In traditional multi-model analysis approaches, the model uncertainties (epistemic and aleatory) are bundled together. The proposed approach allows incorporation and simultaneous propagation of model uncertainties as well as rigorous separation of these uncertainties so that they retain their epistemic or aleatory nature. In addition, these uncertainties are represented by the variation among and within the simulated data ECDFs; the epistemic uncertainties were depicted by an individual ECDF, while the aleatory uncertainties were represented by the dispersion within the given ECDF (Roy and Oberkampf 2011).

Another distinction between the traditional performance measures, such as RMSE, and the area metric is that the proposed approach acknowledges that the development of an exact model to represent the groundwater system is improbable, given uncertainty. Typically, the area metric is used for validation assessment of models that generate entirely probabilistic outputs that can be depicted as distributions (Ferson et al. 2008). Groundwater models are inherently deterministic models that are based on principles of conservation of mass, momentum, and energy (Konikow 1996). However, the explicit incorporation of uncertainty in the multiple model approach and use of multiple sets of observed data creates distributions akin to the probabilistic outputs hitherto used in area metric studies.

The value of the area metric is dependent on the differences between the whole ECDFs of the observed and the simulated data. The model ECDF of a given model is a probabilistic representation derived from a set of deterministic values generated by simulating that model multiple times to represent different system states; these deterministic outputs for each state that are then then arranged into model ECDFs. Also, the observed data from different system states are collated into observed data ECDFs. The differences towards the tails of the distribution also

affect the value of the area metric, as well as the differences in the lower-order moments such as the mean and variances of true distributions. An ECDF accommodates the details about the system behavior that include “typical” behavior as well as the tails. Therefore, a distribution-based comparison facilitated by the area metric approach can be more informative than a comparison of means (Ferson et al. 2008). This avoids overfitting to a particular system state. Instead, it assesses the extent of a broader agreement with the collation of states.

Chapter 3

Methods

Overview

The objective of this chapter is to describe in detail the study area, the modeling technique, and the area metric method. Firstly, the details of the hydrogeologic characteristics of the Brookhaven landfill site and its vicinity was discussed. Secondly, the theory, construction, and simulation of the MODLFOW model was described in detail. Third, the discussion included details of the features included in the model, their classification (either fixed or variable), and their states; the states were fixed in case of fixed features, while the states varied in case of the variable feature class. Fourth, the chapter described the area metric method in detail. This section described the theory of ECDFs and of that of the area metric. This was followed by the description of the application of the area metric method to the case study model in three steps.

3.1. Study Area

The objective of this section is to describe the hydrogeologic characteristics of the Brookhaven landfill site and its vicinity in detail and to highlight the complexity, heterogeneity, and the diversity of interpretations that exist with regard to these characteristics.

The Town of Brookhaven Waste Management Facility is located in the hamlet of Brookhaven, Suffolk County, New York (Figure 3.1). It is bounded by Horseblock Road to the north, Sunrise Highway to the south, the Horizon Village residential community to the west, and Yaphank Avenue to the east (Figure 3.2). The landfill mounds occupy about 180 acres of the 536 acre Facility. Other facilities such as a Material Recycling Facility (MRF), a landfill gas-to-energy recovery system, a waste transfer station, and a Stop-Throwing-Out-Pollutants (STOP) facility also operate at the Facility. Four recharge basins are located on the facility: two to the south of the landfill, one to the east, and one to the north (Dvirka and Bartilucci 2001).

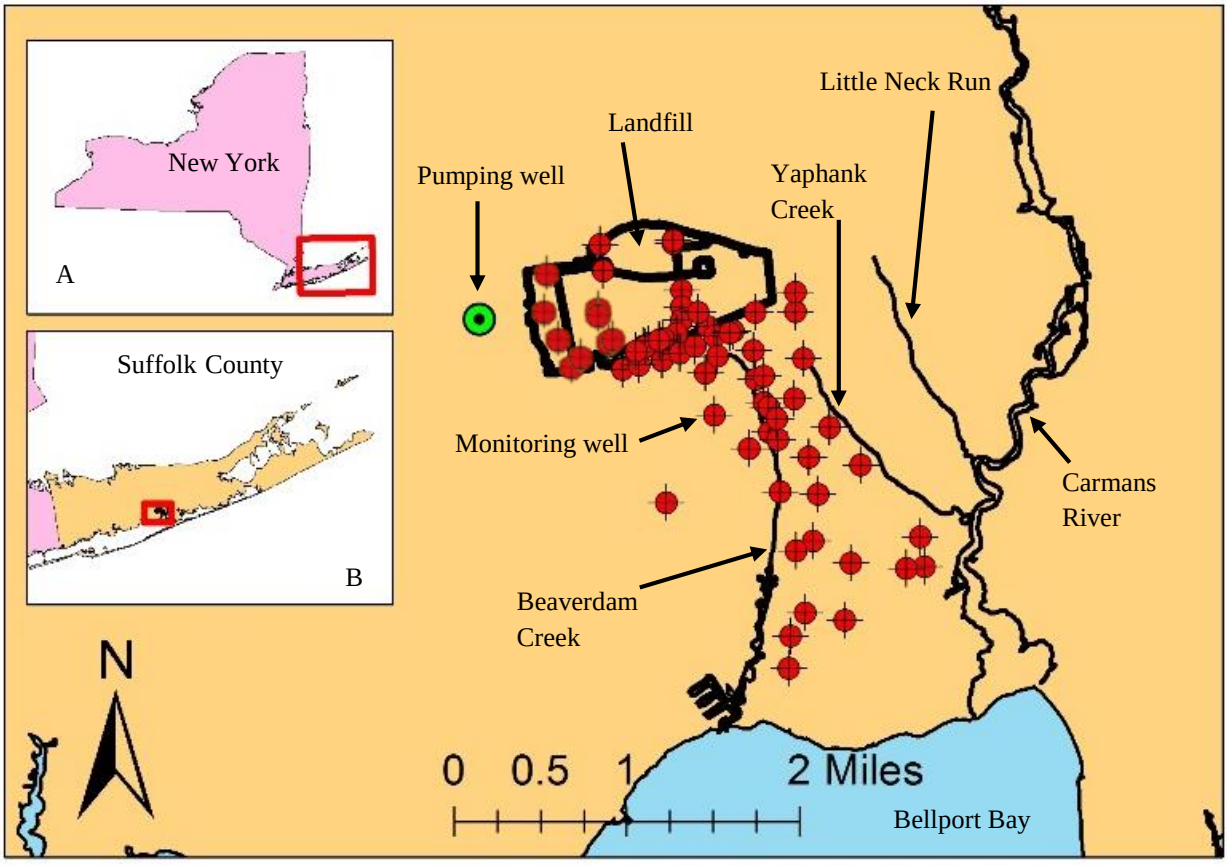


Figure 3.1: Town of Brookhaven West Management Facility, New York state (inset A), and Suffolk County, NY (inset B) (red square indicates area of detail)

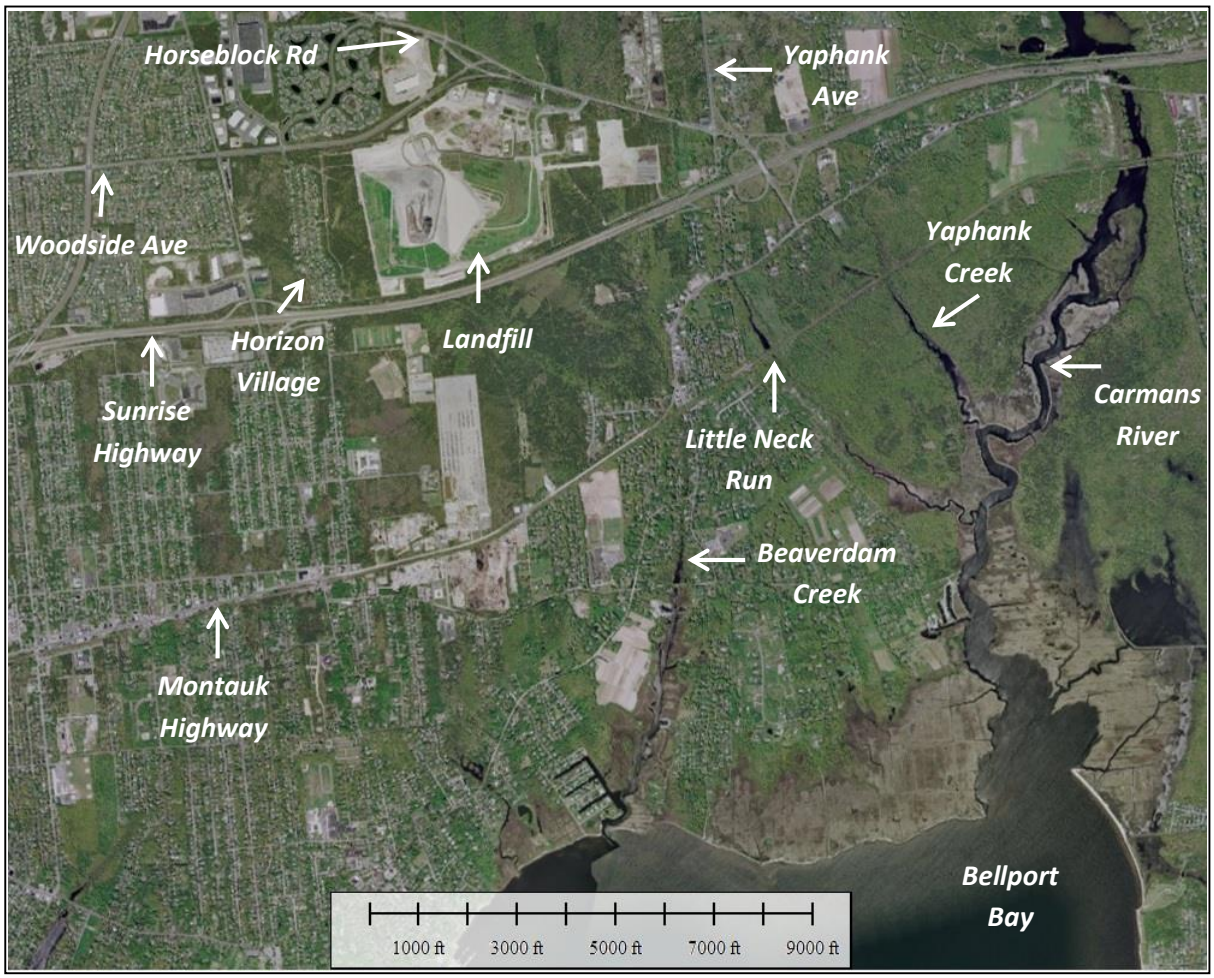


Figure 3.2: Aerial view of the Brookhaven landfill site and its vicinity

The landfill consists of six "cells" that formerly created a general appearance of two landfill mounds. The mound to the east is composed of older cells of the landfill: Cell 1, 2, 3 and 4 (Figure 3.3). The mound to the west is Cell 5. The two mounds were separated from each other by a valley. Cell 6 is constructed in the valley and extends north along Cell 5 and Cell 4. Cells 1, 2 and 3 received municipal solid waste (MSW). Cell 4 received a combination of MSW, construction and demolition (C&D) debris, and incinerator ash. Cell 5 and Cell 6 are restricted to incinerator ash, C&D, and other relatively inert material. All the cells are lined, with liners of varying composition and design.

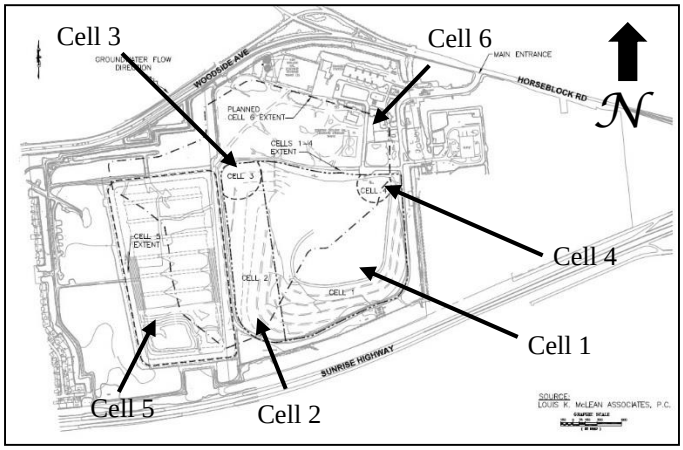


Figure 3.3: Brookhaven landfill site plan (Dvirka and Bartilucci 2011)

The landfill site is located south of the Ronkonkoma moraine on a relatively flat, featureless, and southward sloping outwash plain with gently rolling topography. The topography at the landfill site strikes in the northwest direction and dips in southeast direction. The elevation of the landfill vicinity ranges from a high of 80 feet to the northwest of the site to near sea level to the southeast, near to Great South Bay. Maximum elevation in the area of the landfill is about 250 feet msl, the elevation of the landfill mounds (as of 2009) (Figure 3.4-a). The topographical elevation dips on the western boundary of the landfill site due to presence of the Carmans River valley that is approximately 2 ½ miles wide. The elevation of the valley is noticeably lower than the surrounding area; 20-25 feet msl on the edges while in the center of the valley it lowers to 15-20 feet msl (Figure 3.4-b). The elevation of the valley gently dips towards the Great South Bay where it approaches sea level.

The landfill was excavated below the natural surface elevation into vadose zone sediments, which are predominantly Pleistocene glacial outwash. The bottom elevation of Cell 1 is about 32 feet msl. The basal depth of Cell 2 is unknown, but is assumed to be approximately the same. The bottom elevations of Cell 3 and Cell 4 are about 31.5 feet msl and 39.5 feet msl respectively (Dan Johnson, personal communication, April 6, 2012).

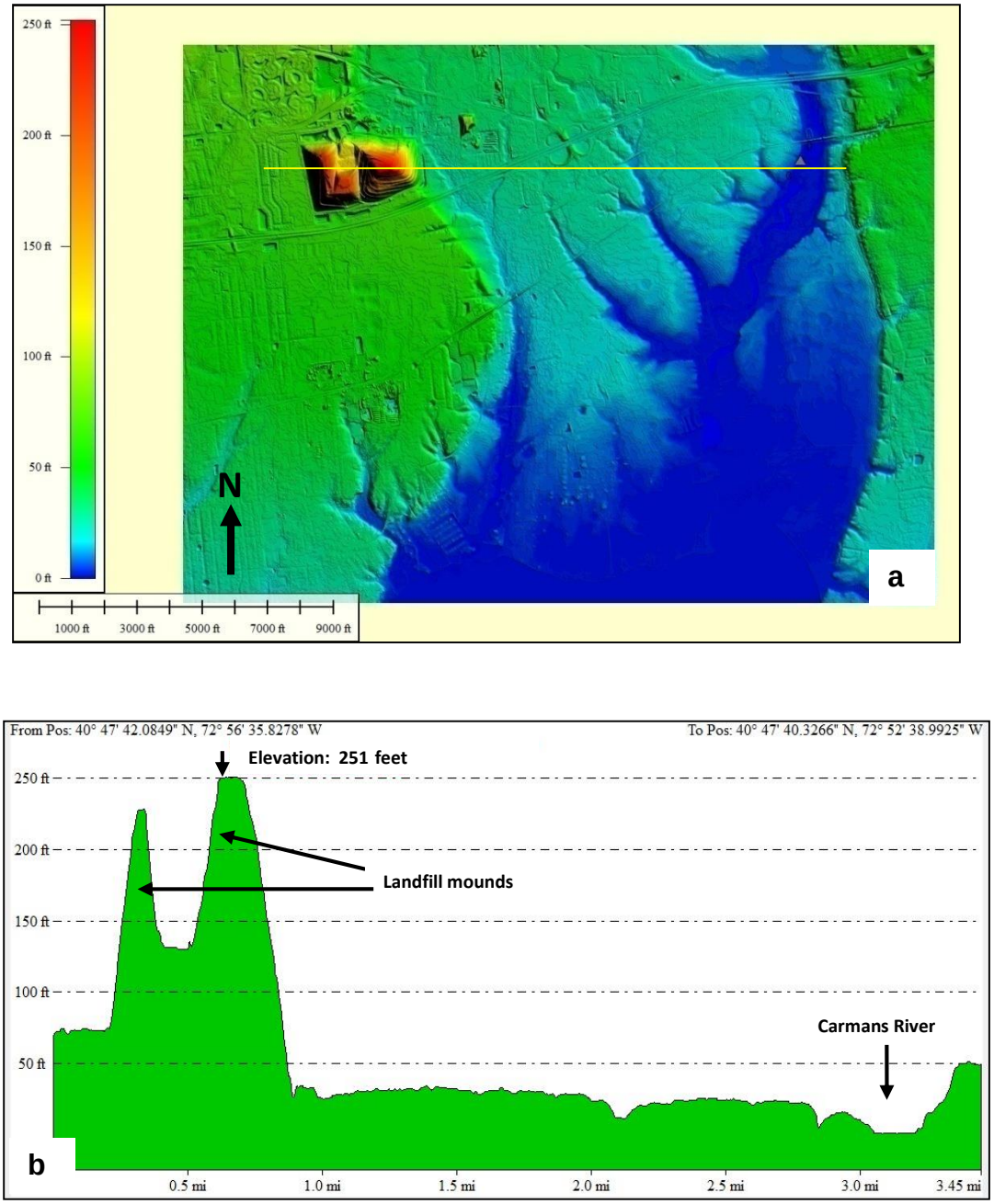


Figure 3.4: Digital Elevation Model (DEM) of the Brookhaven landfill site and its vicinity (as of 2009), (a) plane view, and (b) elevation profile at the elevation profile transect (yellow line)

3.1.1. Geology

The sedimentary units underneath the landfill are of Cretaceous and Pleistocene age (DeLaguna 1963). The deposition of sedimentary units and of the topographical features is a result of movement and structuring of the sedimentary mass brought first by the erosion of northeastern highlands during the upper Cretaceous and thereafter by glacial advances during the late Pleistocene epoch about 100,000 to 18,000 years before present (BP) (Sirkin 1982).These sediments rest on a bedrock surface that dips in the southeasterly direction. The thickness of the sedimentary deposits vary across Long Island from being absent in northwestern Queens County to about 2,000 feet underneath the barrier islands located in the southeastern Suffolk County (Smolensky et al. 1989).

The sedimentary units that overlie the bedrock are (from bottom to top) (i) the Lloyd sand member of the Raritan Formation, (ii) the clay member of the Raritan Formation (Raritan Clay), (iii) the Matawan Group-Magothy Formation, (iv) the Gardiners Clay (with the Monmouth Greensand), (v) the Upper Glacial aquifer (UGA), and (vi) the Holocene or recent deposits (Figure 3.5). The members of the Raritan Formation and the Matawan Group-Magothy Formation are Cretaceous in origin; Gardiners Clay and the Upper Glacial deposits are of Pleistocene age, while the recent deposits belong to the Holocene age (McClymonds and Franke 1972). The Upper Glacial deposits, the Matawan Group-Magothy Formation, and the Lloyd sand member of the Raritan Formation act as aquifers, while the Gardiners Clay (and the Monmouth Greensand) and the Raritan Clay act as confining units that confine the Magothy aquifer and the Lloyd aquifer respectively. The sedimentary deposits along the south shore of Long Island extend beyond the mainland and the barrier beaches out into the Atlantic Ocean (Scorca et al. 1995).

Thirty years of study at the site report that the contamination of the local groundwater from the landfill leachate is restricted to the Upper Glacial aquifer (Dvirka and Bartilucci 2011). Therefore, the vertical extent of the model domain included three hydrogeologic units (from bottom to top): shallow portions of the Magothy aquifer, the Gardiners Clay and the Monmouth Greensand, and the Upper Glacial aquifer along with the Holocene deposits.

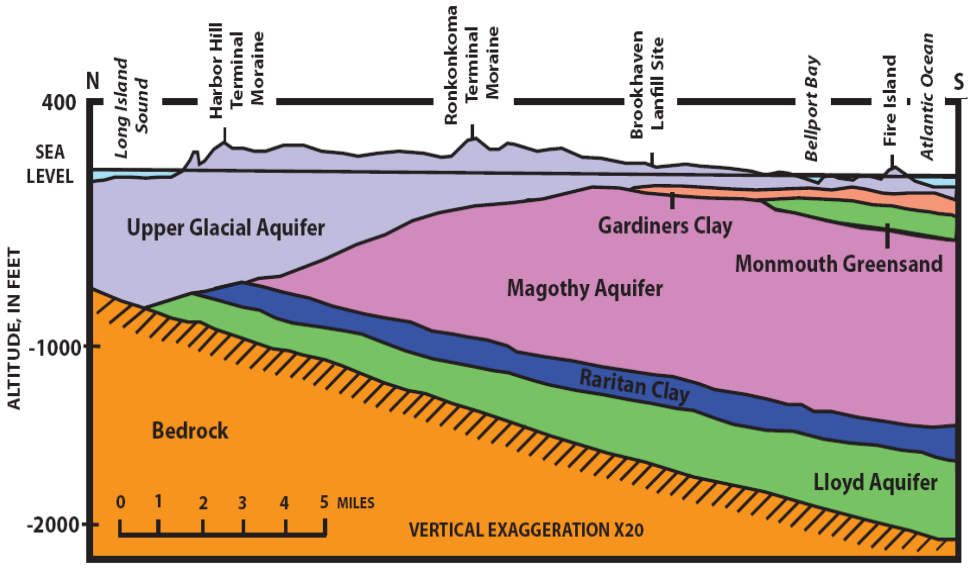


Figure 3.5: Generalized cross section of Long Island geology (modified from McClymonds and Franke 1972)

3.1.1.1. Magothy Aquifer

The Matawan Group-Magothy Formation (undifferentiated) is the sedimentary deposit of the late Cretaceous period. The Matawan Group-Magothy Formation unconformably overlies the Raritan Clay and is unconformably overlain by either the Monmouth Greensand or by the transitional deposits of Pleistocene age (Aronson et al. 1983). Geophysical investigations suggest that the upper surface of the Matawan Group-Magothy Formation begins between -110 feet to -130 feet msl. The unit is about 900 feet thick at the landfill site (Dvirka and Bartilucci 1994a).

The Matawan Group-Magothy Formation is composed of fine to medium quartzose sand interbedded with silt, gray clay with abundant amounts of lignite, pyrite, marcasite, organic matter, and clay (Dvirka and Bartilucci 1994b; Lindner and Reilly 1983; Aronson et al. 1983). Geologic borings indicate the presence of localized clay lenses (discontinuous zones of solid clay of variable thickness) (Smolensky and Feldman 1992).These clay lenses can be up to 50 feet thick; their thickness usually grows with the growing thickness of the Magothy aquifer.

The hydrologic name of the Matawan Group-Magothy Formation is the Magothy aquifer. The Magothy aquifer is considered a principal aquifer on Long Island (Aronson et al. 1983; DeLaguna 1963). The shallower sections of the Magothy aquifer are comprised of fine sand and therefore have very low hydraulic conductivity (K) value and aquifer potential. The average

horizontal hydraulic conductivity (K_x) of the aquifer as a whole was estimated to be 54 feet/day with a range from 27 feet/day to 134 feet/day (McClymonds and Franke 1972). Other estimates of K_x include 67 feet/day (Soren 1971), 268 feet/day (Isbister 1962) for Nassau and Queens Counties, and 50 feet/day (Franke and Cohen 1972). The mean K_x was calculated to be 1 feet/day (Bouwer-Rice rising head test) and 0.033 feet/ ay (Hazen method) based on hydrogeologic investigation at a shallow Magothy well MW11-M (220 feet below grade or about -150 feet msl) at the landfill. Based on the average K_x value of 1 feet/day, the average groundwater velocity for the shallow Magothy aquifer was calculated to be about 0.0043 feet/day (Dvirka and Bartilucci 1994a). Also, the vertical hydraulic conductivity (K_z) based on the same test was estimated to be 0.03 feet/day (Dvirka and Bartilucci 1994a, 1996a). The anisotropy ratio estimated to range from 30:1 to 100:1 (Smolensky et al. 1989; Lindner and Reilly 1983; Franke and Cohen 1972); the ratio was estimated to be 33:1 at the landfill site (Dvirka and Bartilucci 1994a).

3.1.1.2. Potentially Semi-confining Unit (PSU)

The low-permeability material found underneath the landfill is referred to in this report as the "potentially semi-confining unit" (PSU). This unit is not a classified geologic unit, but rather to be taken as a hypothetical layer that is an ensemble of the low permeability units found in the study area, including the Gardiners Clay and the Monmouth Greensand, where present. This unit provides semi-confining to confining conditions that may hydraulically disconnect groundwater flow between the Upper Glacial and the Magothy aquifer.

Gardiners Clay is composed of greenish-gray to gray clay, with medium to coarse quartzose sand, and is interbedded with silt, mixed layer clays, and fine gravel (Wexler 1988a; Wexler and Maus 1988; Voorhis 1986). The presence of glauconite, a green, iron silicate mineral of the mica group, is responsible for the greenish appearance of the clay unit (Koszalka 1984). The Gardiners Clay is generally believed to be of marine origin, although it has also been described as a brackish water, lagoonal, non-marine cold water, or pro-glacial deposit (Sirkin 1986). A biostratigraphic investigation (Stone and Borns 1986) indicated the age of the clay unit to be about Sangamon interglaciation period (~38,000 years BP).

The PSU tapers landwards and thickens seaward. The thickness of the Gardiners Clay ranges from 0 feet at its northern limit to about 90 feet beneath the barrier islands (Doriski and

Wilde-Katz 1983); it may extend beyond the south shore beaches. The elevation of the unit ranges from -100 to -150 feet msl and up to -200 feet msl beyond the barrier beaches (Smolensky et al. 1989; DeLaguna 1963).

The Gardiners Clay can act as a natural barrier to leachate flow and its presence underneath the leaky landfill can help prevent the contaminated groundwater in the Upper Glacial aquifer from mixing with deeper waters. Also, groundwater in the Upper Glacial aquifer is unlikely to percolate deeper into the Magothy aquifer because of the horizontal and vertical components of the local groundwater flow regime. Further, the shallow Magothy is composed of low permeability fine sand that can lead to generation of anisotropic conditions that resist downward movement of groundwater. Therefore, only the shallow section of the Magothy aquifer is simulated in the model. The title "Gardiners Clay" is not to be taken literally; although it connotes that the unit is composed of solid clay, the geologic evidence suggests that the composition of the unit is highly variable, from sandy to silty to solid clay. In addition, the characteristic greenish clay appears only in certain borings; clays of other colorations, such as brown, white, black, and gray are also found at similar depths at several boring locations, including sand hardpan in one instance. Different opinions exist about the local elevation and thickness of the Gardiners Clay underneath the landfill site and its vicinity (Dvirka and Bartilucci 1994a; Buxton and Modica 1992; Smolensky et al. 1989; Fanning, Phillips and Molnar 1986; Voorhis 1986; Gerathy and Miller 1985; DeLaguna 1963; Weiss 1954). Therefore, the position, thickness, and extent of the Gardiners Clay in the vicinity of the landfill is uncertain.

The confining abilities of the Gardiners Clay are further enhanced if it is underlain by another low permeability unit, the Monmouth Greensand. Monmouth Greensand has hydrologic characteristics similar to that of the Gardiners Clay. Therefore, these units are treated as a combined hydrogeologic unit: the Potentially Semi-confining Unit (PSU).

3.1.1.3. Upper Glacial Aquifer (UGA)

The Upper Glacial deposit is the uppermost geologic unit of the Pleistocene age. The Upper Glacial deposits are composed of stratified, tan to brown, coarse to fine grained sand and gravel with a small amount of clay and silt (Perlmutter and Gerathy 1963). The sand is mostly quartzose and contains alkali feldspar, mica, amphibole, biotite, chlorite, and hornblende (DeLaguna 1963; Perlmutter and Gerathy 1963). Generally, the coarseness of the sand increases

from the bottom to the top; however, the basic lithology of the unit remains the same (Dvirka and Bartilucci 2001). Under the landfill, the bottom 15 feet to 20 feet of the deposit was found to be made up of reddish brown to brown, fine, micaceous, silty sand (Wexler 1988a).

The Upper Glacial deposits form the uppermost, water table aquifer on Long Island – the Upper Glacial aquifer (UGA). The thickness of the saturated portion of the aquifer ranges from 30 feet to 120 feet (Perlmutter and Gerathy 1963); the range is 90-135 feet at the landfill (Dvirka and Bartilucci 1994a). The deposits in the upper sections are generally coarse and readily yield water, while the deeper sections have better sorted sands with lower permeability. The K_x values range between 607 feet/day (Bouwer-Rice Rising Head Test) and 30 feet/day (Bouwer-Rice Rising Head Test) for the shallow section, between 208 feet/day (Hazen method) to 58 feet/day (Bouwer-Rice Rising Head Test) for the intermediate section, and from 363 feet/day (Bouwer-Rice Rising Head Test) to 17 feet/day (Bouwer-Rice Rising Head Test) for the deeper sections of the UGA. The overall average K_x is between 200 feet/day to 300 feet/day (Lindner and Reilly 1983). Anisotropy ratio ranges from 2:1 to 24:1 with an average ratio of 10:1 (Smolensky et al. 1989; Gerathy and Miller 1985; Lindner and Reilly 1983; Reilly et al. 1983). Based on the anisotropy ratio of 10:1 and the K_x of 270 feet/day, K_z can be calculated to be 27 feet/day. The porosity (η) of the UGA averages about 0.33 (McClymonds and Franke 1972), while the effective porosity values (η_e) range from 0.25 to 0.30 (Gureghian et al. 1981; Kimmel and Braids 1980). The mean transmissivity value was estimated to be 310,000 gallons per day/feet, while the mean specific yield for the UGA was estimated to be 0.22 (Lockwood, Kessler and Bartlett (LKB), 1994). Based on the average K_x value of 270 feet/day (K), porosity of 0.30 (n), and the horizontal hydraulic gradient for the shallow UGA (i) of 0.001 feet/feet, the average linear groundwater velocity was calculated to be about 1 feet/day (365 feet/year) for the UGA (using Darcy's law, $v = K*i/n$; Fetter 2001).

3.1.1.4. Holocene Deposits

Holocene deposits are the youngest deposits on Long Island. These deposits overlie unconformably on the Upper Glacial deposits and are deposited and reworked either by wind or wave action, or by human activity. The Holocene deposits are not considered hydrologically important because of their thinness and localized, limited deposition (DeLaguna 1963).

3.1.2. Hydrology

The climate on Long Island is generally mild and humid, and it is influenced by the westerly patterns that drive continental weather systems eastward. Extreme diurnal temperatures are moderated due to bordering water bodies such as the Atlantic Ocean and Long Island Sound. Long Island is occasionally affected by dramatic coastal weather systems such as the "nor-easters" and hurricanes (Peterson 1987). The average annual temperature is about 55ºF (Climate Report for Islip, NY, National Weather Service, January 2013).

Precipitation events are distributed fairly evenly throughout the year on Long Island, primarily from fronts sweeping west. Major rainfall events occur during the winter season and are often associated with coastal storms that generate northeasterly winds. The largest storms during warm periods (June to November) are associated with coastal storm systems from the south, although these tend to be frontal systems (Peterson 1987). Winter precipitation may convert into snowfall under favorable air temperatures, although snow or sleet accounts for less than 10 percent of the total precipitation (Koszalka 1984).

Variable accountings of Long Island precipitation exist: 44 inches/year (Miller and Fredrick 1969); 43 inches/year for Suffolk County (Krulikas 1986); 44.7 inches/year. at the Islip MacArthur Airport (National Weather Service, New York, NY, Climate Report for Islip, NY, collected January 2, 2013); 48.5 inches/year at the National Weather Service weather station at Upton (about 6.5 miles north of the landfill); and 47 inches/year around the landfill (Wexler and Maus 1988). The long term average annual precipitation, to the period 1949-2013, is 48.3 inches per year from Brookhaven National Laboratory (Figure 3.6).

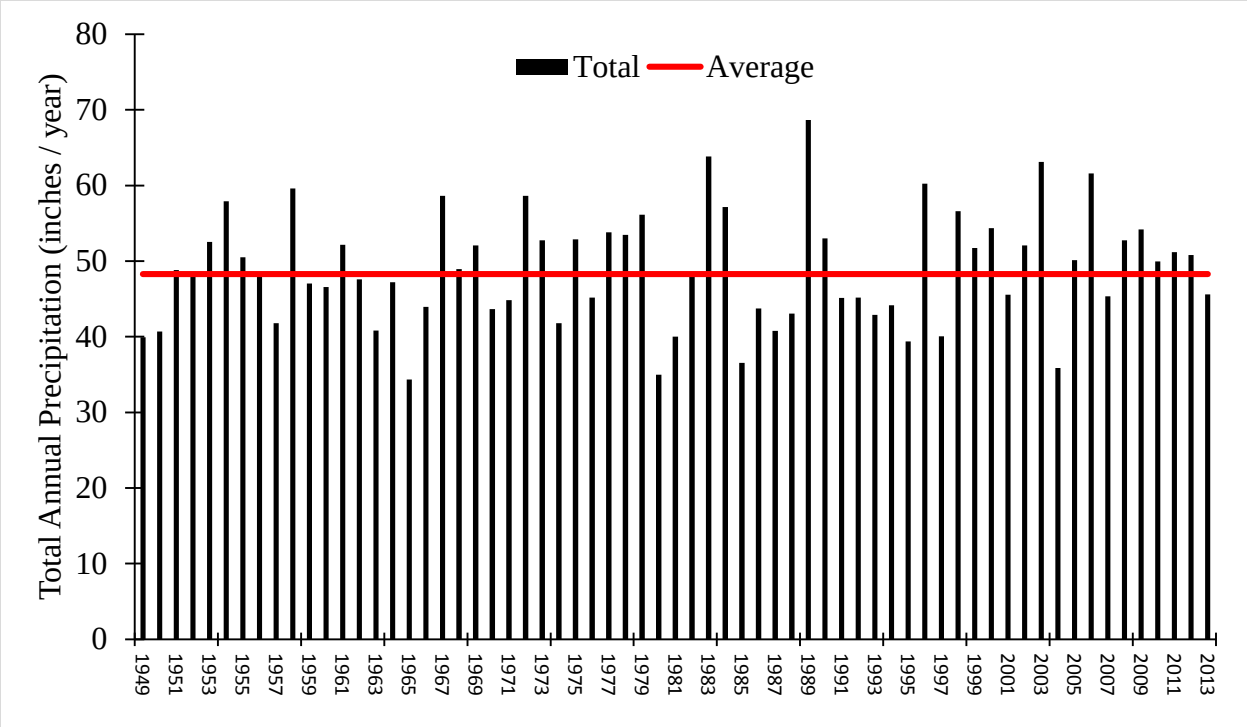


Figure 3.6: Annual Precipitation at Upton, NY (1949 to 2013) (average = 48.3 inches, in red) (Source: http://www.bnl.gov/weather/4cast/MonthlyPrecip.htm)

3.1.2.1. Recharge

Precipitation that falls on ground first passes through the root zone of plants and the unsaturated soil zone. Once the soil moisture deficiency of the unsaturated zone (vadose zone) is replenished, the remainder of the infiltrated water percolates to the saturated zone by gravitational flow. The flat topography of Long Island, along with highly permeable nature of the outwash deposits, promotes ready subsurface percolation of precipitation (Spinello and Simmons 1992). Long Island has flat topography and no long-duration snow pack. The groundwater system on Long Island is isolated from continental systems. Therefore, recharge to the groundwater system occurs only from local, immediate precipitation (Garber 1986; Koszalka 1984; McClymonds and Franke 1972). The natural recharge is affected by the amount of precipitation, as modified by evapotranspiration and runoff (Peterson 1987):

Natural Recharge = Precipitation – (Evapotranspiration + Direct Runoff)

Recharge to the water table aquifer (here, the UGA) occurs directly from precipitation. Recharge to deeper, confined aquifers occurs only in the Deep Recharge Zone: a zone

surrounding the band across the middle of Long Island that forms the north-south groundwater divide (Koppelman 1978). Recharge to the deeper, confined aquifers occurs because of vertical flow through the UGA. The amount of recharge that reaches the deeper aquifers depends on (i) the amount of precipitation in the vicinity of the Deep Recharge zone, (ii) presence of confining layer(s); (iii) flow patterns in the UGA, and (iv) the saturation level in the overlying soils.

Recharge is generally greatest during the cold season (October-March) (late fall, winter, and early spring) because lower temperatures and shorter daylight periods lead to dormant vegetation and little evaporation; frozen ground impedes recharge, however. Recharge is usually lowest during the warm season (April-September) (late spring-summer-early fall) because warmer temperatures, more daylight, and more vegetation growth result in greater evapotranspiration (Eckhardt and Wexler 1986). Recharge may be greater or lesser than expected during any one particular year due to variations in rainfall, temperature, and plant growth. Variations in precipitation usually dominate changes in recharge, so that dry winters have little recharge and very wet summers result in larger recharge.

Roads and other development, landfill liners and mounds, and presence of recharge basins redistribute recharge locally (Pearsall and Aufderheide 1995). Storm water management practices on Long Island lead to increased recharge because recharge basins and catch basins used to manage human-caused run-off inhibit evapotranspiration (Seaburn and Aronson 1974).

On average, the annual recharge value is estimated to be half of the annual rainfall. Steenhuis et al. (1985) estimated the recharge percentage to be 50.7 % of annual rainfall in Mineola, 54.6 % in Patchogue, and 54.1 % in Setauket for the period 1968-1975. Wexler and Maus (1988) estimated annual precipitation around Brookhaven landfill to be 47.4 inches and used 24.6 inches/year as a recharge value. Peterson (1987) modeled recharge based on precipitation patterns and soil types, and estimated recharge near the landfill to be 22-23 inches/year.

3.1.2.2. Direct Run-Off

Under natural conditions, the amount of precipitation entrained in surface runoff for most of Long Island is negligible due to flat topography, highly permeable soils, and vegetation cover. Increases in urbanization lead to more impervious surfaces such as roads, parking lots, and roofs,

land-clearing, and, devegetation that, in turn, reduces recharge and increases runoff. Direct runoff was estimated to be 1 percent of the total rainfall in the vicinity of the landfill (Wexler 1988a). Runoff percentages may be greater near Sunrise Highway and other roads, especially on steeper slopes. Long Island uses recharge basins and catch basins to manage most run-off. These structures increase water deliveries to the subsurface and so may elevate recharge rates above natural levels (Seaburn and Aronson 1974).

3.1.2.3. Evapotranspiration

The term evapotranspiration encompasses two phenomena: (i) the physical process of evaporation from exposed water and moist soil; and, (ii) the biological process of transpiration through plants as they take water (the nutrient carrier) from roots and release it through leaves (Peterson 1987). Evapotranspiration is highest on Long Island during the warm season (April-September) because of warmer temperatures, more light, and more vegetation growth (Warren et al. 1968). During the warm season, recharge can be non-existent if all infiltrated water is taken up by plants or evaporates directly (Busciolano et al. 1998).

The depth of the root zone, the height of the water table, and the ability of the soil to hold moisture (its field capacity) also determine the amount of evapotranspiration. Direct transpiration from the aquifer by plants is possible if the water table is within 4 feet of the ground surface (Pluhowski and Kantrowitz 1964).

The outwash and gravel base for soil found in southern and central Long Island results in soils with less field capacity compared to soils based on less permeable tills, which are more common on the north shore. In general, most Long Island soil has low field capacity (Warner et al. 1975). The evapotranspiration fraction of total precipitation varies from 21.2 inches (46.6 %) near Bridgehampton where soil is a sandy loam type with shallow root vegetation to 26.8 inches (57.9 %) near Setauket where the soil is a sandy loam type soil with mature forest (Peterson 1987). The predominant natural soils at the landfill site are Plymouth sandy loam and Riverhead sandy loam (about 40% each) with some of Carver and Plymouth sands (10 % each) (Warner et al. 1975). Peterson (1987) estimated evapotranspiration to be 22 inches/year in the vicinity of the landfill. No direct experiments were conducted to estimate the level of evapotranspiration at the landfill.

3.1.2.4. Salt-water Bodies

Bellport Bay, part of the lagoonal, estuarine Great South Bay in the South Shore Estuary system, is located about 2.5 miles south of the landfill. Bellport Bay is bounded by Fire Island to the south, Brookhaven hamlet to the north, and the Mastic-Shirley peninsula to the east. All streams in the study area discharge into the Bellport Bay.

3.1.2.5. Streams

Streams on Long Island are almost entirely fed and sustained by groundwater discharge because the contribution of groundwater, or baseflow, is about 90 to 95 percent of the total flow of the streams under natural (pre-development) conditions (Spinello and Simmons 1992; Wexler and Maus 1988; Prince et al. 1988; Peterson 1987; Pluhowski and Kantrowitz 1964). In other words, Long Island streams are essentially groundwater drains. Stream flow during dry weather spells depends on the groundwater levels adjacent to the stream. When groundwater heads are high, and where the ground surface is lower than the water table elevation, groundwater seeps through the streambed resulting in stream flow. Conversely, when the water table drops below the streambed elevation, seepage reverses and the stream dries (Gerathy and Miller 1985; Prince 1980; Pluhowski and Kantrowitz 1964).

Fresh water streams near the landfill site include Beaverdam Creek, Yaphank Creek, Little Neck Run, and Carmans River. Beaverdam Creek lies to the eastern edge and Carmans River to the western edge of the valley, while Yaphank Creek and the Little Neck Run traverses through the valley in a southeastern direction. Beaverdam Creek is closest; its headwaters are found south of Sunrise Highway, immediately southeast of the landfill. Carmans River is the largest stream near the landfill, located approximately 1¾ miles to the east of the landfill. Yaphank Creek and Little Neck Run are tributaries to Carmans River and are located approximately ¾ miles southeast of the landfill (Figure 3.7).

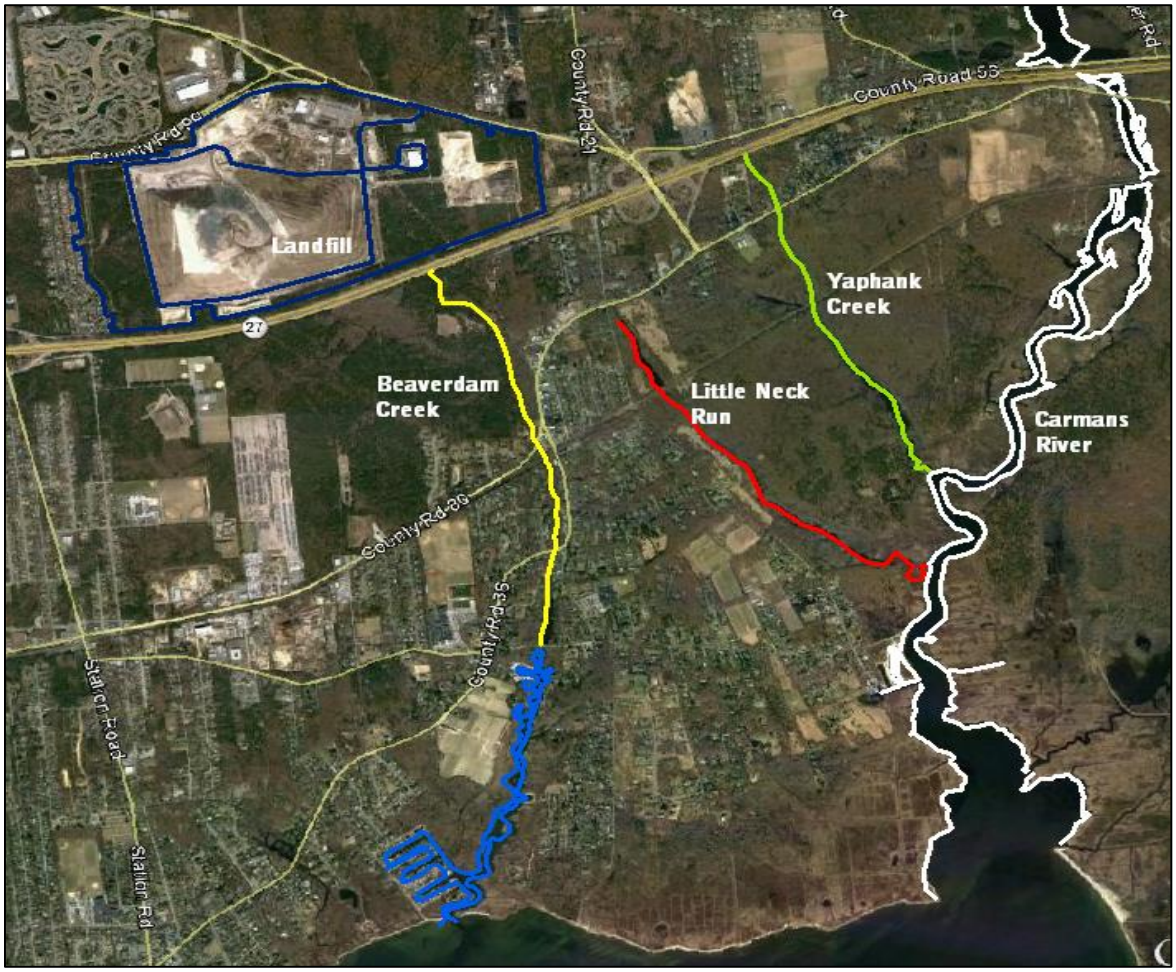


Figure 3.7: Areal image showing the non-tidal portion of Beaverdam Creek (in yellow), the tidal portion of Beaverdam Creek (in blue), Little Neck Run (in red), in Yaphank Creek (in green), and Carmans River (in white)

Beaverdam Creek is closest to the landfill. The headwaters of Beaverdam Creek are near Sunrise Highway, while it discharges into Bellport Bay in the south. Beaverdam Creek is approximately 2.5 miles long and is marine (tidal) for about 1.1 miles upstream from Bellport Bay to Beaverdam Road (Figure 3.8-a). The Creek is ditched in its freshwater portion in a number of places for mosquito control and also has been dredged in its tidal section to support boating. The average baseflow for the Creek was observed to be 1.35 cubic feet/second (ft^3s^{-1}) near the intersection of Montauk Highway and South County Road (Wexler 1988a). Dvirka and Bartilucci (2012) measured an average 2.43 ft^3s^{-1} at approximately the same location from May 2011 to March 2012 (Table 3.1). The hydraulic gradient along Beaverdam Creek varies. The

gradient is relatively flat (0.002 feet/feet) between the headwaters and sampling location BD-4 (approximately 1,300 feet north of the gaging station), then it steepens (0.004 feet/feet) between BD-4 and BD-2 (about 770 feet south of the gaging station) and then it decreases again further downgradient due to the flat topography (Dvirka and Bartilucci 1994a).

Period	05/05/11	05/27/11	06/12/11	07/15/11	09/08/11	16/11/11	01/18/12	03/07/12
Flow	2.65	4.58	1.62	1.50	3.71	1.71	2.47	2.27

Table 3.1: Stream flow measurement (flow in ft^3s^{-1})

Figure 3.8: (a) Aerial View of Beaverdam Creek: non-tidal (red) and tidal sections (white) and (b) Photograph of the Non-tidal Section of Beaverdam Creek. The yellow star indicates the approximate position of the gaging station for flow measurements made by Wexler (1988a) and Dvirka and Bartilucci (2012).

Carmans River lies east of the landfill (Figure 3.9). The river is 11 miles long, extending from Cathedral Pines County Park in Middle Island to Bellport Bay. The width of the stream varies from 3 to 50 feet, while the depth varies from a few inches to about 6-8 feet. The Carmans River basin is relatively undeveloped. The salt and brackish tidal marshes along Carmans River south of Montauk Highway constitute large sections of the Wertheim National Wildlife Refuge. A weir on the river at the Southaven County Park north of Sunrise Highway maintains a lake about 6 feet above the tidal river (Cashin Associates 2002). Waters downstream of the weir are considered to be tidal (Wexler 1988a). The Carmans River is the natural hydrologic divide for the regional shallow groundwater sub-system that flows in southeasterly direction. The USGS estimated an annual discharge rate of 37.8 ft^3s^{-1} at an upgradient location near the Long Island Expressway. The average annual baseflow for the Carmans River was estimated to be 56 ft^3s^{-1} (Wexler and Maus 1988) at a recording station located at the Victory Avenue dam. The fresh water flow rate increases to 72 ft^3s^{-1} at the mouth of the river (Cashin Associates 2002).

Carmans River has two tributaries to its west: Yaphank Creek and Little Neck Run (Figure 3.9). Flow in Little Neck Run begins south of Montauk Highway (at the railroad bridge) although stagnant pools of water are found north of the railroad. There is a perennial flow in Yaphank Creek north of Montauk Highway, but it does not extend north beyond Sunrise Highway. Both streams are less than a mile in length and become tidal about 1,000 feet south of the railroad tracks (Wexler 1988a). The average baseflow was estimated to be 0.1 ft^3s^{-1} for Little Neck Run and 0.12 ft^3s^{-1} Yaphank Creek, at stations at the railroad tracks (Wexler and Maus 1988).

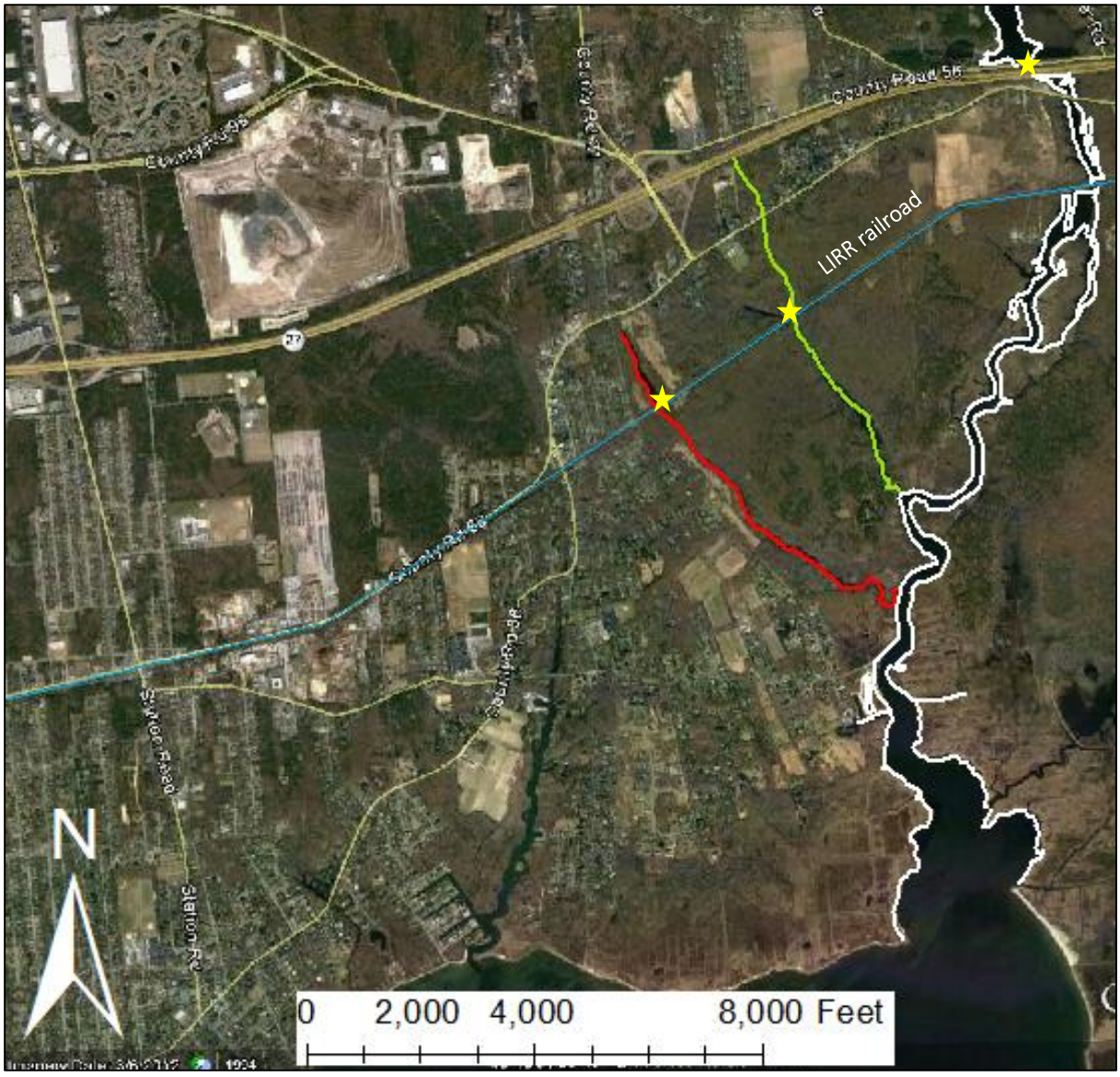


Figure 3.9: Little Neck Run (red), Yaphank Creek (green), and Carmans Rivers (white). The yellow stars show the approximate locations of flow measurements made by Wexler (1988a).

3.1.2.6. Consumptive Use of Water

Consumptive use of water occurs when water is drawn from a system and is not returned to the system after use. There is little consumptive use of groundwater in the vicinity of the landfill. Public water was supplied to most houses in the study area around 1990 due to concerns regarding potential effects of the landfill plume on downgradient private drinking water wells. The Suffolk County Water Authority (SCWA) has a public water supply well field along Bellport Road west of the landfill, but the Water Authority has an interconnected system.

Therefore, it is not certain that water supplied with public water comes from that well site. Therefore, there may be some water imported into the area; but also some water may be exported to other Suffolk County residents. There are no sewer systems in the study area; all houses and businesses use subsurface disposal systems (septic systems or cesspools) for sanitary wastewater. This means that if water in the public supply system comes from outside of the study area, the sanitary systems are net producers of recharge to the system.

Houses and businesses west and east of the public water supply area use private wells for water supply. These wells are typically installed in the Upper Glacial aquifer, 40 to 60 feet below the water table. The use of subsurface sanitary waste treatment means there is negligible consumptive use of this water (Spinello and Simmons 1992; Buxton and Reilly 1985). There are two farms in the study area that use irrigation during times where plant water demands exceed soil moisture availability. All of these wells used to be in the Upper Glacial aquifer. Because the landfill plume reached the Hamlet Organic Garden (H.O.G.) Farm, the Town installed two wells into the Magothy aquifer in 2010 to provide irrigation water. Because of evapotranspiration and export of agricultural products from the area, these withdrawals can be considered to be minor consumptive uses of local water. There are no major industrial water uses in the study area.

3.1.2.7. Groundwater

Groundwater levels in the UGA respond to changes in precipitation and climatic conditions; generally, the groundwater levels rise as precipitation increases, and fall as precipitation amounts decrease. Pressure responses to changes in groundwater table elevations are usually rapid (returning to equilibrium in several days) because of the good hydraulic connectivity throughout the UGA, but large precipitation events may result in temporary downward vertical gradients (Wexler 1988a). The response rate is affected by variations in evapotranspiration, ground saturation levels, runoff associated with snowmelt or storm events, and the water transmission from upgradient to the downgradient. All of these factors are reflected in the groundwater level measurements (Aphale and Tonjes 2010).

Groundwater levels in the Town of Brookhaven range from a maximum of slightly less than 100 feet msl near the center of the Town to near mean sea level near the coastline. In the vicinity of the landfill, the water table elevation generally ranges from 3 feet (or less) msl (at or near the shoreline) to 30 feet msl. The depth to the water table from the natural ground surface

ranges from a little more than 50 feet in the northwest to 0 near streams. The saturated thickness of the UGA ranges from 90 to 135 feet (Dvirka and Bartilucci 1994c).

The direction of horizontal flow in the UGA, the underlying Gardiners Clay, and upper sections of the Magothy aquifer near the landfill is southeasterly (Dvirka and Bartilucci 2001; Eckhardt and Wexler 1986) (Figure 3.10). Groundwater in the UGA flows horizontally with little downward flow, except as driven by recharge inputs, with very little discharge into the Magothy aquifer (Wexler 1988a). A mapping of equipotential heads based on local water table aquifer measurements shows general agreement with the regional flow map. Local differences include obvious discharge of the aquifer into Beaverdam Creek and Carmans River.

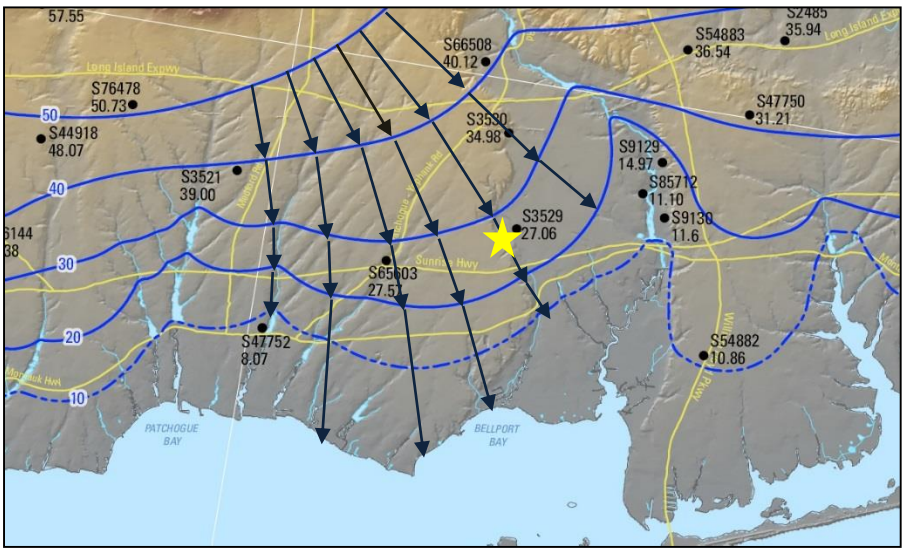


Figure 3.10: Potentiometric altitude of the UGA (blue lines). Arrows show approximate horizontal direction of the groundwater flow (modified from Monti and Busciolano 2009). Star indicates approximate location of the landfill.

The coarser sediment composition of the UGA makes it more conductive than the Magothy aquifer. Consequently, the rate of movement of groundwater in these aquifers also varies; water in the UGA generally moves faster compared to the Magothy aquifer or the Lloyd aquifer. Also, the head gradient in the UGA is steeper than the Magothy aquifer. Potentiometric pressure in the center of Long island is greater in the UGA than in the deeper aquifers, creating the potential for recharge of the deeper aquifers from the UGA. The head pressure declines quickly in the UGA becoming essentially zero at the shore line. This reverses the relationship and the potentiomentric surface of the underlying Magothy aquifer can exceed the UGA. This

creates potential for the vertical flows of groundwater from the underlying deeper aquifer into the UGA (Koppelman 1978). The presence of the PSU can retard the flow of groundwater between aquifers and prevent equalization of the pressure differences (Aphale and Tonjes 2013).

There is a potential for flow from the UGA into the Magothy aquifer north and west of the landfill (Tonjes and Wetjen 2002). Well pairings in the center of the south perimeter of the landfill describe a transition between downward and upward potential flows. Another well pairing north of Montauk Highway in Brookhaven hamlet where the deepest well is screened in the confining unit, has slightly greater head in the deepest well compared to Upper Glacial wells at the same locations indicating a potential for upward flows. Well pairs south of Montauk Highway in Brookhaven hamlet have much greater head in the Magothy aquifer compared to the UGA. Figure 3.11 shows a "transition zone" (Dvirka and Bartilucci 2001): a change from horizontal flow to upward flow in the Magothy aquifer.

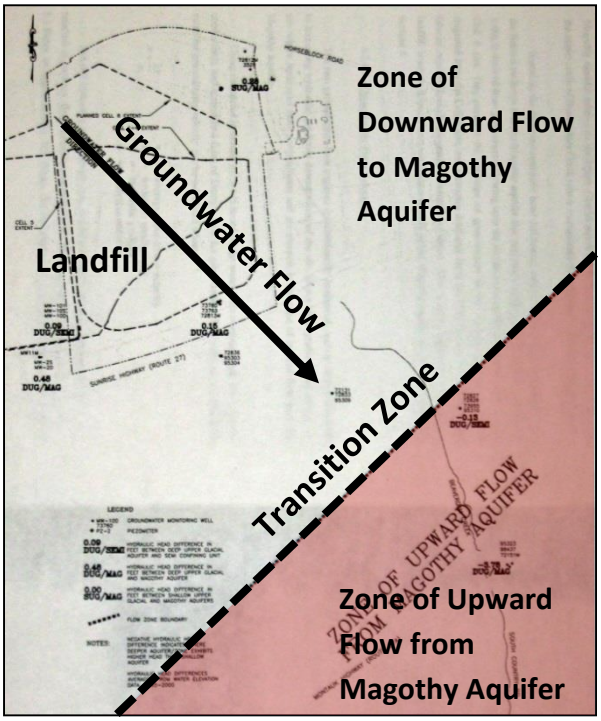


Figure 3.11: Transition Zone (Dvirka and Bartilucci 2001)

Groundwater head measurements have been collected in the vicinity of the landfill for 30 years (1981-2010) by a number of agencies and organizations. USGS and Suffolk County Department of Health Services have several wells that are monitored regularly. USGS (during

the cooperative agreement), Town consultants, and Stony Brook researchers have monitored much larger networks of wells at generally more irregular intervals. These data have all been collated by Stony Brook University on behalf of the Town.

3.1.2.8. Groundwater Data Used in this Study

The observation data used in this study consists of measurements of groundwater heads made at a distributed network of 133 wells located in the vicinity of the landfill (Figure 3.12). The location, the screen depth, and the aquifer in which these wells are screened differ; but all of these wells are screened in either the UGA or in the Magothy aquifer (Figure 3.13). For example, well S3529 is screened in the shallow UGA at 45 feet below the ground surface (about - 8 feet msl), while well S72812 is screened at 164 feet below the ground surface (about -157 feet msl). The location maps indicate that majority of the wells are located downgradient of the landfill site and screened in the UGA. Thus, the representation of the groundwater condition in the model domain is biased because of the bias in the three dimensional spatial arrangement of the wells.

Some wells are set in a well cluster; these wells share the same location but different screen depths. For example, a cluster of three wells – MW5-S, MW5-I, and MW5-D – is located at the southeastern edge of the landfill property; well MW5-S is screened at 15 feet msl, well MW5-I is screened at -13 feet msl, and well MW5-D is screened at -82 feet msl. Suffixes – -S, -I, and -D – denote shallow, intermediate, and deep screens. 40 wells were screened in the shallow UGA (from top surface to -40 feet msl), 62 wells were screened in the intermediate UGA (-40 feet to -70 feet msl), and 25 wells were screened in the deep UGA (-70 feet to -100 feet msl, or the top of PSU or Magothy). Four wells – 72812, 72151M, 72813, MW11M – are screened in the shallow Magothy aquifer (the suffix M indicates that the well is screened in the shallow Magothy aquifer). Two wells – 95310 and 96202 – are possibly screened in the PSU (Figure 3.13).

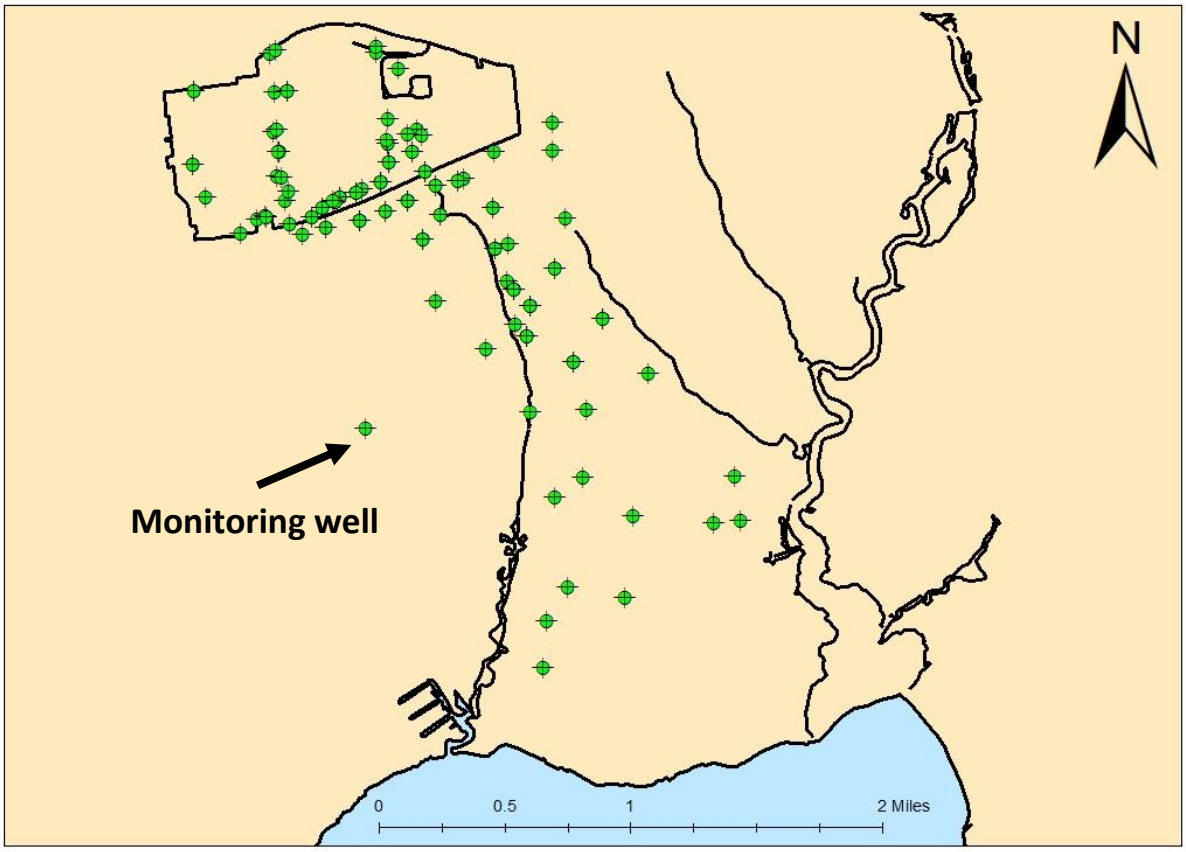


Figure 3.12: Locations of wells used in the study

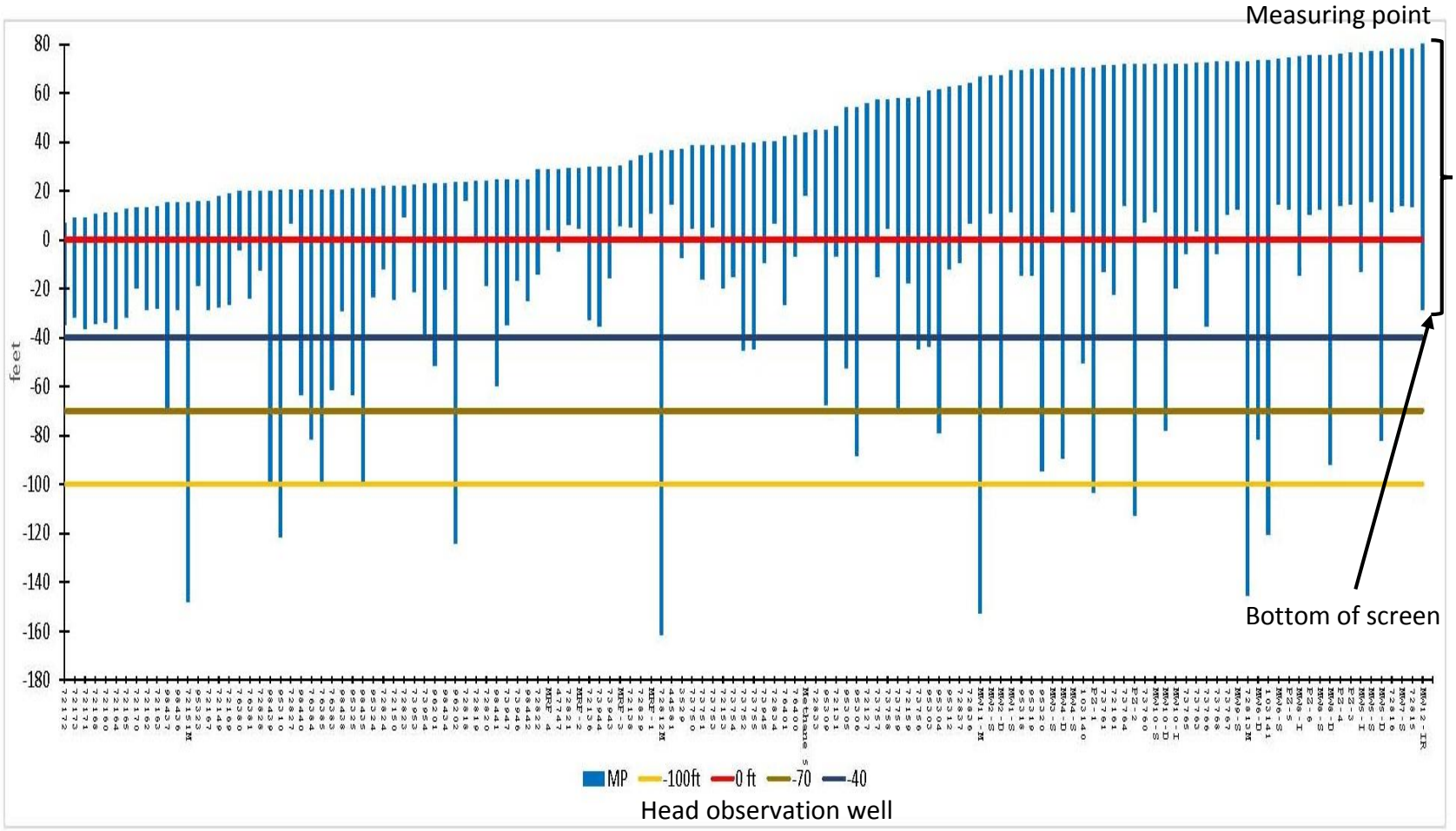


Figure 3.13: Profile view of the screen zones of wells

3.2. Model Used

3.2.1. Groundwater Flow Simulation Model for the Brookhaven landfill

The present groundwater flow simulation model serves three main purposes.

First, the primary objective of the USGS model was to identify factors that affect the contaminant transport in the groundwater in the surroundings of the Brookhaven landfill. The groundwater flow simulation model USGS developed in the second report formed the basis for the contaminant transport model described in the third report. In tune with the USGS's objective and approach, the present modeling exercise establishes a groundwater flow simulation model that would then form the basis to develop the contaminant transport modeling solution.

Second, the landfill is located south of the regional ground water divide and the direction of ground water flow in its vicinity is southeasterly. The conductive nature of the principal aquifers in the region, the Upper Glacial and Magothy aquifers, results in mostly horizontal ground water flows with a small and locally noticeable vertical component. Hence, the advective-dispersive propagation of contaminants through the sub-surface units would be a function of the direction, rate, and volume of groundwater flow through different hydrogeologic units. Therefore, the present modeling effort focuses on achieving a reasonably accurate interpretation of the hydrogeologic regime of the study area ("getting the flow right"). Confining units, such as the potentially semi-confining unit (PSU), create partial hydraulic disconnect downgradient of the landfill; however, the presence and the configuration of the PSU is uncertain. Therefore, multiple model conceptualizations were developed and tested for their representativeness.

Third, the landfill model is being reconstructed almost 25 years after the work done earlier by Wexler published in 1988. Since 1988, a number of investigators have studied the regional and site-specific hydrogeologic properties of the area. These newly available data, such as geologic maps and cross sections, and geophysical/ geotechnical boring and well logs could help better define the geologic framework of the study area. In addition, more data are available pertaining to water quantity and quality; and, arguably modeling practices have evolved for the better due to advances in computational power and graphic abilities of computers. The present model incorporated model features additional and different than the USGS's model – three dimensional model (as opposed to USGS's 2-D model), finite difference grid (as opposed to

USGS's finite element grid), inclusion of the shallow Magothy and the potentially semi-confining units in addition to the Upper Glacial aquifer (as opposed to USGS model where the flow was simulated only in the Upper Glacial aquifer). In addition, the landfill has expanded its footprint resulting in reconfiguration of the local landscape and changes in the inventory of the observational data points (for example, destruction of head observation wells and the construction of new wells). These changes also called for an updated modeling effort. Hence, the present model simulates the groundwater flow around the Brookhaven landfill by incorporating select additional/updated data.

In summary, the present modeling effort aims to assist the Town of Brookhaven's efforts to address the landfill leachate issue. The present model marks the continuation of the modeling efforts carried out by the USGS about 25 years ago. The present model incorporates additional and updated data to better understand and configure the study area. The present model focuses on the groundwater flow in the study area due to the likelihood of transportation of leachate through groundwater that, in turn, could contaminate deeper, more pure groundwater sources.

3.2.2. MODFLOW

One very common modeling code is MODFLOW (Modular Flow) (McDonald and Harbaugh 1984). MODFLOW uses a three-dimensional finite difference governing equation to simulate the groundwater flow, combining Darcy's Law and the principle of conservation of mass (Anderson and Woessner 1992, p. 15; McDonald and Harbaugh 1988) (Equation 3.1):

$$\frac{\partial}{\partial x}\left(K_x\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_y\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_z\frac{\partial h}{\partial z}\right)+W=S_s\frac{\partial h}{\partial t} \quad (3.1)$$

where,

K_x, K_y, K_z = values of hydraulic conductivity along the x, y, and z coordinate axes, that are assumed to be parallel to the axis of hydraulic conductivity (L^1T^{-1})

S_s = specific storage of the porous material (L^{-1})

h = potentiometric head (L^1)

W = volumetric flux per unit volume representing sources and/or sinks of water, with W <0 for flow out of groundwater system, and W >0 for recharge (T^{-1})

t = time (T^1)

∂h= partial derivative of hydraulic head with respect to 3 partial derivatives of space (∂x, ∂y, ∂z) and time (∂t) (L^1).

MODFLOW provides an algebraic solution to a finite difference numerical approximation. The groundwater head at any given node ($h_{i,j}$) in the mesh-centered finite difference grid is calculated as an average of the groundwater heads from four of the nearest neighbor nodes ($h_{i-1,j}$, $h_{i+1,j}$, $h_{i-1,j-1}$, $h_{i,j+1}$) (Equation 3.2). The solution is calculated for all the nodes using a five-point operator that moves a "star" of five points across all the nodes in the grid in a systematic fashion (Wang and Anderson 1982, p. 25) (Figure 3.14).

$$h_{i,j}=\frac{h_{i-1,j}+h_{i+1,j}+h_{i,j-1}+h_{i,j+1}}{4} \quad (3.2)$$

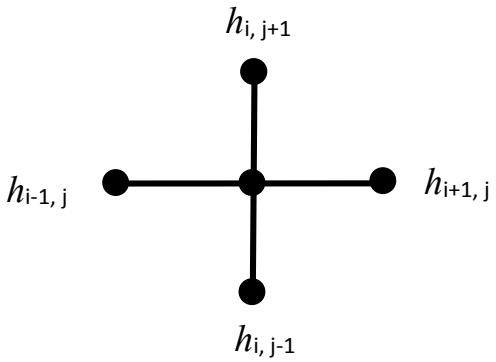


Figure 3.14: The five-point operator

The two-dimensional finite difference approximation is given as follows (Wang and Anderson 1982, p. 69) (Equation 3.3):

$$\frac{h^{n}_{i+1,j} - 2h^{n}_{i,j} + h^{n}_{i-1,j}}{(\Delta x)^2} + \frac{h^{n}_{i,j+1} - 2h^{n}_{i,j} + h^{n}_{i,j-1}}{(\Delta y)^2} = \frac{S}{T}\,\frac{h^{n+1}_{i,j} - h^{n}_{i,j}}{\Delta t}\,\frac{R^{n}_{i,j}}{T} \tag{3.3}$$

where,

$\Delta x = \Delta y$ = length of grid (L^1)

Δt = length of the time-step (T^1)

n = time level (T^1); $n-1$ = time level one step earlier; $n+1$ = time level one step later

$R^{n}_{i,j}$= recharge per unit time per unit aquifer area at point (i,j) (L^3T^{-1})

S = storage coefficient (dimensionless)

T = transmissivity coefficient (dimensionless)

MODFLOW generates an iterative solution to the finite difference numerical approximation of the groundwater flow equations. The solution to a groundwater flow problem begins with an initial estimate of groundwater head; this initial value is iteratively adjusted until a pre-established error criterion comparing model solutions is satisfied, or a number of cycles determined by the modeler by trial-and-error depending on the rate of convergence to the final solution. A variety of iterative schemes are used: Gauss-Sidel iteration, Successive Overrelaxation (SOR), Jacobi iteration (Wang and Anderson 1982, p. 25), and PCG2 (Hill 1990).

MODFLOW uses a combination of different packages classified either as formulators and solvers of finite difference equations. The modular nature of the MODFLOW program allows the modeler to call specific packages as they are needed. This streamlines modeling

exercise because packages that are not required for construction and simulation of a model are not activated. This also allows the modeler to have a better control and flexibility.

New packages have been developed for specific purposes and can be coupled with the MODFLOW code. For instance, MODPATH was developed for the purpose of particle tracking and visualization (Pollock 1994). PEST (Parameter Estimation) is an example of an automated parameter estimation program that selects values for parameters that minimize the difference between the observed and the simulated values ("the model error") (Doherty and Hunt 2010). UCODE (Poeter et al. 2005) is another commonly used parameter estimation program. ZONEBUDGET calculates a sub-regional water budget for the model using cell-by-cell flow data (Harbaugh 1990). The MT3DMS package simulates the fate and transport of contaminants, including their advection, dispersion, and chemical reactions (Zheng 2010).

Upgrades of the original MODFLOW have been developed including MODFLOW-1996 (Harbaugh and McDonald 1996), MODFLOW-2000 (Hill et al. 2000; Harbaugh et al. 2000), and MODFLOW-2005 (Harbaugh 2005). A number of graphical user interfaces are available such as Visual MODFLOW (Waterloo Hydrogeologic, Inc. 2006), ModelMuse (Winston 2009), Groundwater Modeling System (GMS) (Aquavevo, LLC 2010), and Groundwater Vistas (Environmental Solutions Inc. 2015). Post-processers are used to generate and view output from the model code.

Visual MODFLOW v. 4.2 (Waterloo Hydrogeologic, Inc. 2006), a groundwater modeling platform based on MODFLOW-2000, was used to develop the three-dimensional, steady-state, finite difference groundwater flow simulation models for the landfill. MODFLOW-2000 code is a well-tested, peer-reviewed code for groundwater model simulation and hence model code verification was not made part of the exercise. The initial head – or the guess of the head values at beginning of the model simulation – was fixed at 30 feet. The PCG2 solver package was used for the simulation; the specifications of the model solution are given in Table 3.2. Time to complete one model run was about 4.2 seconds for each iteration.

Numerical Engine	MODFLOW 2000
Max outer iterations	25
Max inner iterations	30
Pre-conditioning method	Cholesky (NPCOND=1) or Polynomial (NPCOND=2)
Head-change criterion	0.01
Residual criterion	0.01
Relaxation parameter	1
Upper bound of estimate	Fixed estimate (NPBOL=2)
Damping factor	1

Table 3.2: PCG2 solver specifications

3.2.3. Fixed Model Variables and States

The model features with uncertainty incorporated into the model are termed recognized errors, while those uncertain features not incorporated into the model are termed unrecognizable errors (Oberkampf et al. 2002). The number of model features that go into a complex, physically distributed model such as the present model can be very large and each feature is accompanied by its own type, level, and nature of uncertainty. A pragmatic approach was adopted where a select set of model uncertainties were acknowledged and represented using multiple model configurations so that the number of models representing these uncertainties will be limited. In the present study, uncertainties associated with eight variable features were incorporated into the model (section 3.2.3), and the remainder of model features were kept fixed. The variable features were represented by different values or states – different structural configurations or conditions – in different model iterations. On the other hand, the fixed features, as their name indicates, had a fixed value or state in all models. The following discussion narrates the model settings as well as the fixed features of the model.

3.2.4. Model Domain

The model domain covered about 32.5 mi^2, of which about 23.2 mi^2 portion was active and 9.4 mi^2 was inactive. It occupied from 0 to 2.79016E4 feet in the X direction and 0 to 3.2427E4 feet in the Y direction. The eastern boundary of the model was defined by Carmans River, the southern boundary was defined by the Great South Bay. The western boundary –

Swan River – was simulated by a general head boundary (GHB), while the northern boundary – the hydrologic divide on Long Island – was simulated as a constant-head (CHD) boundary. Both, the GHB and the CHD boundaries represented the net effects of the external natural boundary on the system within the model domain. They allow the simulation of hydraulic boundary condition without the need to extend the model domain to encompass the actual boundary condition.

3.2.4.1. Grid

A variably-spaced finite difference grid of 427 rows and 333 columns was set on the model domain, defining 142,191 cells altogether. The model domain was defined by either by naturally existing hydraulic boundaries or their numerical representation. The grid-cell dimensions were reduced from the model boundaries towards its center; coarsest grid cells (270 feet X 180 feet) were at the periphery of the model domain while the finest grid cells (35 feet X 45 feet) were towards middle of the model where the landfill and the well network were located. The grid was smoothed to maintain the aspect ratio of the grid under the recommended limit of 5 (Anderson and Woessner 1998, p. 69). The principle axis of the grid was aligned southeasterly; in the main axis of the groundwater flow (Figure 3.15).

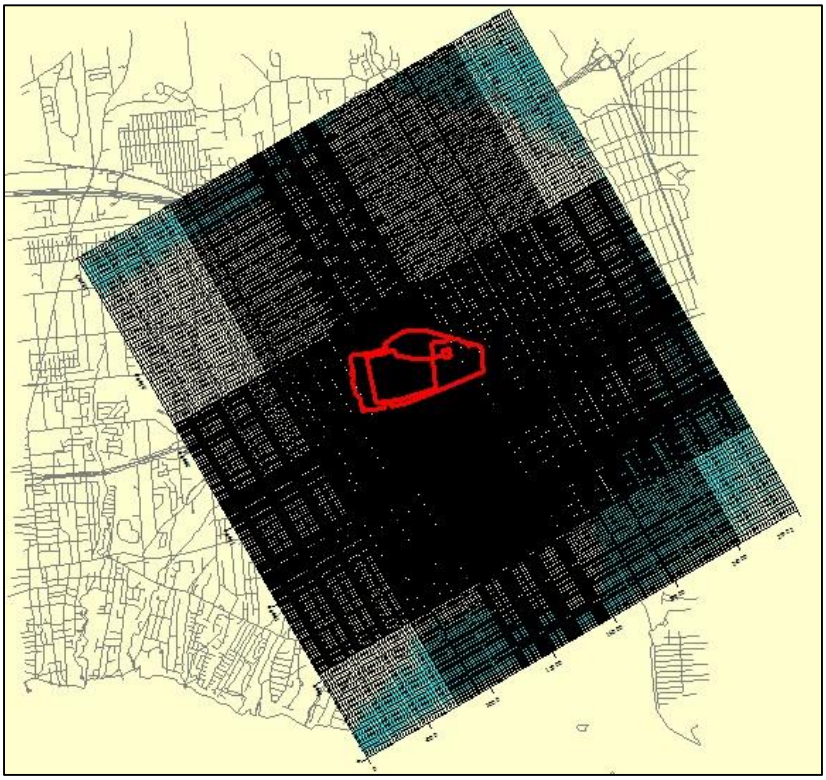


Figure 3.15: Model grid

3.2.4.2. Elevation

The ground surface elevation was imported from the National Elevation Data set (NED) 10 m resolution (http://ned.usgs.gov/index.html). The elevation Bellport Bay was fixed at 0 feet msl; that was considered as the mean sea level (msl). The elevation of the bottom of the landfill was represented by a fixed value of 40 feet msl (Figure 3.16).

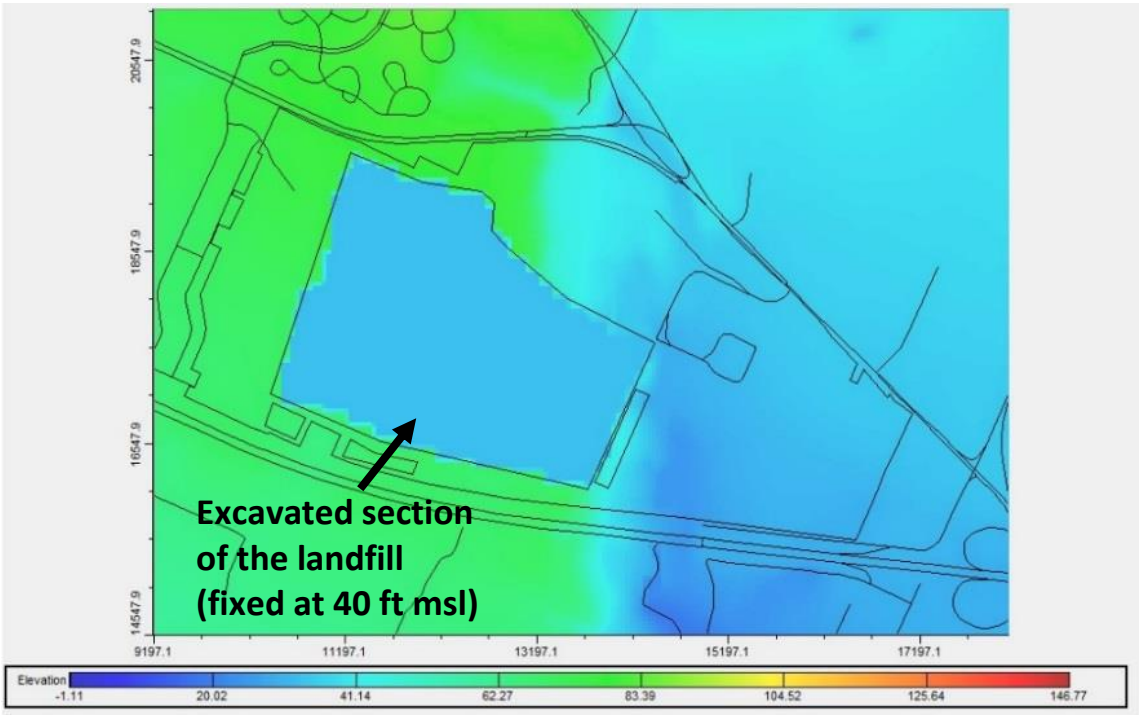


Figure 3.16: Elevation at the landfill mounds (set to the bottom of the excavated portions)

3.2.4.3. Vertical Discretization

The vertical extent of the model domain ranged from the topographic surface of the study area to the shallow reaches of the Magothy aquifer. Given the southeasterly sloping layering of the model, the deepest point of the model domain was about 250 feet below msl. Non-horizontal model layers are useful in representing hydrogeologic units that are sloping because each unit can be assigned to a discrete model layer (Harte 1994). Therefore, the vertical domain of the model comprised of five layers; the upper three layers ("L1", "L2", and "L3") represented the UGA, the fourth layer represented the PSU, and the bottom layer (L5) represented the Magothy aquifer (Figure 3.17). The boundaries between these layers were variable; only the topographic surface and the model bottom remained fixed.

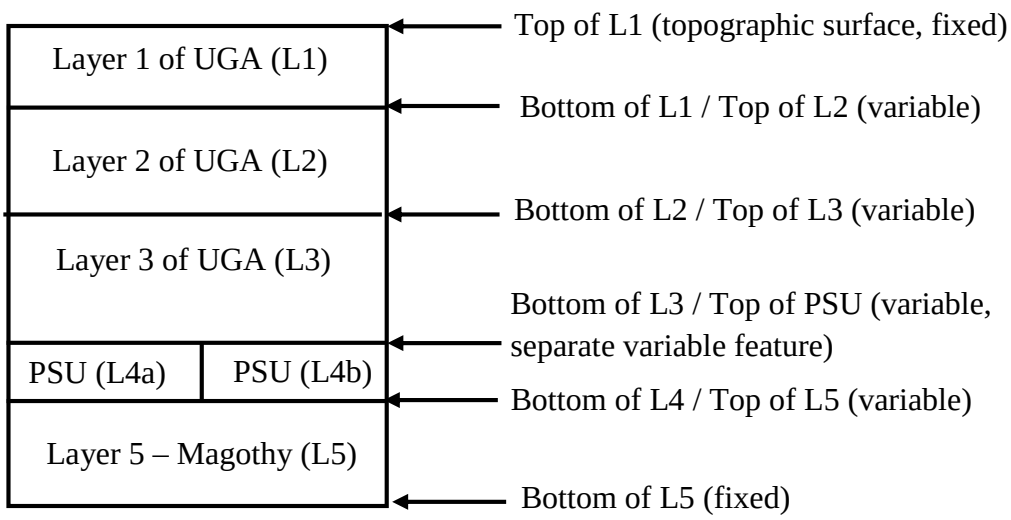


Figure 3.17: Conceptual depiction of the vertical discretization of the model

The UGA shows a tendency of downward fining, that is, the coarseness of the aquifer material progressively decreases from near the surface to the bottom of the aquifer unit. The deposits in the upper sections are generally coarse and readily yield water, while the deeper sections have better sorted sands with lower permeability (Dvirka and Bartilucci 1994a; Lindner and Reilly 1983; McClymonds and Franke 1972). The transition of conductivities is likely to be continual rather than crisp. The modeling program requires crisp transitional boundaries between varying conductivity zones. Increasing the number of zones increases discretization that would estimate the true, continual transition; however, this increases the numerical burden of simulation. Therefore, the UGA was vertically divided into 3 layers – L1, L2, and L3 – to represent a practical, conservative transition where the thickness and the hydraulic conductivities of these layers was varied.

Similarly, the lithological evidence indicate that sediments grouped into the PSU range from sandy silt to solid clay and there appears to be southwardly fining of the sediments (Wexler 1988; Wexler and Maus 1988; Voorhis 1986; Koszalka 1984; Pluhowski and Kantowitz 1964; Perlmutter and Gerathy 1963). Consequently, the conductivity of the materials in the PSU layer also reduces from north to south (Aphale and Tonjes 2013). Here, this phenomenon was represented by grading the hydraulic conductivity of the layer four (L4) into three separate zones of progressively low-hydraulic conductivities.

Zone 1 spreads from the northwestern corner of the model domain to the north of the landfill site in a crescent-like fashion in accordance with the regional arrangement of the Gardiners Clay on Long Island (Figure 3.18). This zone represented the L4a of model domain that indicated the absence of PSU and it had the same K_h as that of the UGA. Zone 2 was depicted as an intermediate zone between the high-permeability Zone 1 and the low-permeability Zone 3. It appeared in the form of a 5,000 feet thick band south of Zone 2. Zone 2 underlie the landfill property. Finally, Zone 3 covered the area south of Zone 2 and extended to the southern limit of the active model domain.

Layer 4 was horizontally divided into two sections: the northern "L4a" and the southern "L4b". L4a constituted Zone 1 and it represented the UGA and simulated the absence of the PSU. L4b constituted Zone 2 and Zone 3 and it represented the PSU.

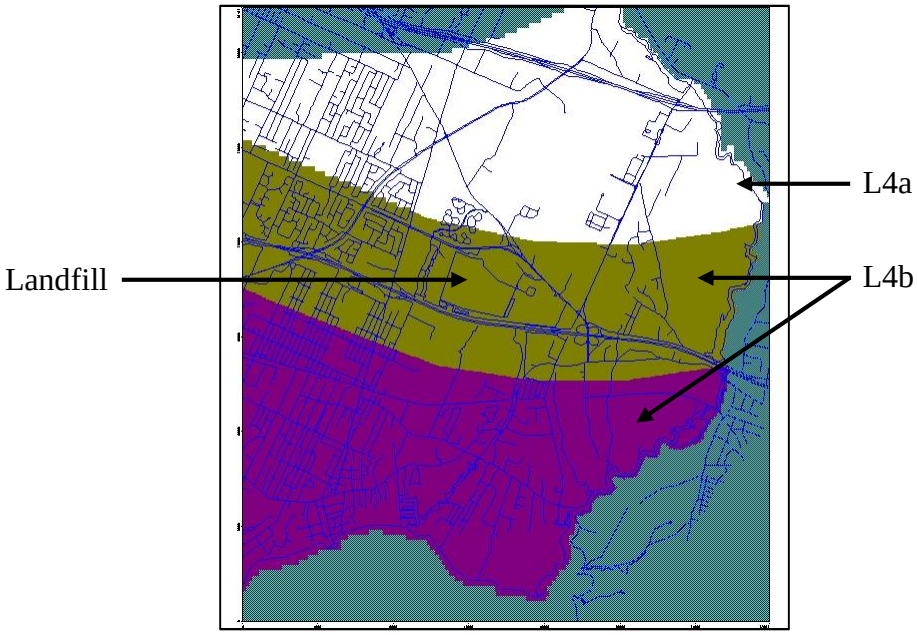


Figure 3.18: Top surface of the showing L4a showing the UGA (white) and L4b showing the PSU (combination of green and magenta); Conductivity zones in L4: Zone 1 (white), Zone 2 (green), Zone 3 (magenta)

3.2.4.4. Inactive Zone

Three portions of the model domain were designated as inactive zones: the portion south of the coastline encompassing the Great South Bay, the portion north of the upper constant head boundary of the UGA, and the portion to the east of Carmans River (Figure 3.19). The cells in the inactive zone did not take part in the model calculations. This arrangement facilitated

cropping portions of the finite-difference rectilinear grid that were hydrogeologically disconnected.

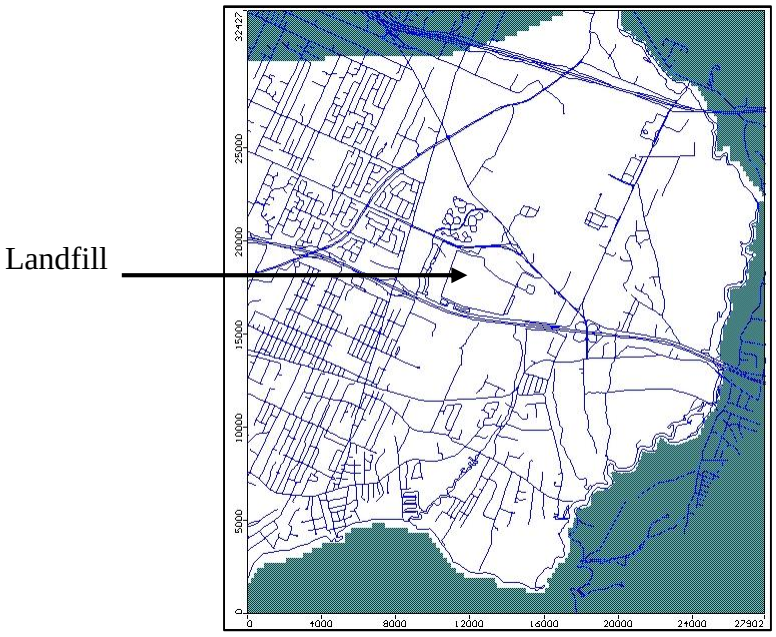


Figure 3.19: Inactive zones (blue)

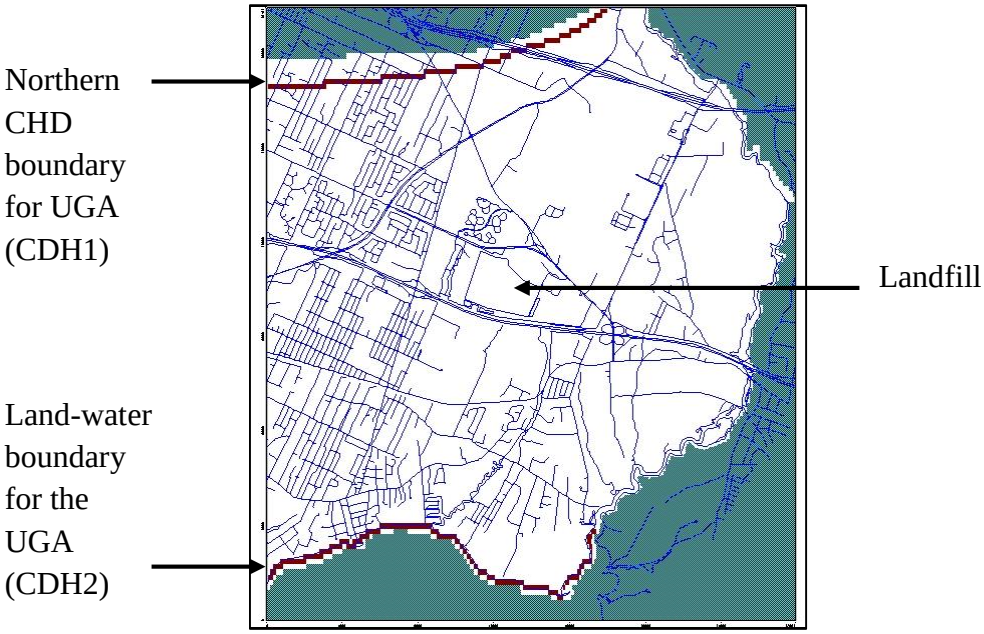
3.2.4.5. Constant Head (CHD) Boundaries

A constant head (CHD) boundary is set in the model domain when the real-world hydrologic boundary is at a considerable distance farther from the study area. CHD represents the Dirichlet boundary condition where a fixed value of groundwater head is maintained throughout the simulation period by moving necessary and unbounded amounts of water in or out of the aquifer (Mercer and Faust 1981). Three locations of CHD boundaries of the model domain were determined from available maps of equipotential contours (Wexler and Maus 1988).

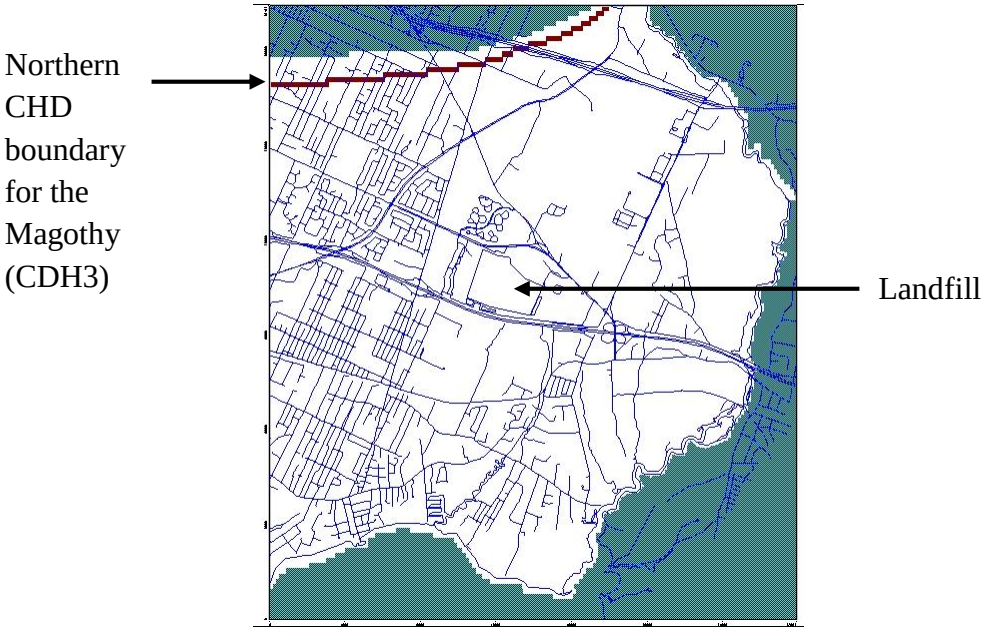
The first CHD boundary was located at the northwestern edge of the model domain in an approximately semi-circular fashion (Figure 3.20-a). This CHD boundary simulated the hydrologic divide located about 3 miles north of the landfill, at approximately the center of Long Island, which runs east to west. This CHD boundary represents an aleatory uncertainty associated with the landfill model; therefore it is technically not a fixed feature.

The second CHD boundary was located in the first layer (L1) and it represented the land-surface salt-water interface between the Town of Brookhaven and the Great South Bay. A constant value of 0 feet was assigned to this boundary (Figure 3.20-b).

The third CHD boundary was located in the fifth layer (L5) at the same location as that of the first CHD boundary. This boundary simulated the groundwater flow from the northern hydrologic divide in the Magothy aquifer. A constant value of 40 feet was assigned to this boundary (Figure 3.20-b).



(a)



(b)

Figure 3.20: CHD boundaries for (a) the L1 of the UGA, (b) L5 – the Magothy aquifer

3.2.4.6. General Head Boundary (GHB)

The general head boundary (GHB) is a Neumann condition where the flux is specified corresponding to a specified hydraulic gradient (Mercer and Faust 1981). Swan River was simulated as a GHB. Swan River is located approximately 14,000 feet west from the western perimeter of the landfill (Figure 3.21-a). It represents the real-world hydrologic boundary to the west of the landfill; the start-of-flow of Swan River is approximately 3,400 feet to the west of the western edge of the model, while the mouth of the stream is approximately 13,000 feet away from the western edge of the model. The GHB boundary lines the western edge of the model domain from approximately 15,000 feet south of the Long Island Expressway in the north stretching up to the coastline in the south (Figure 3.21-b). The GHB boundary was assigned to the first layer (L1). Additional attributes assigned to the GHB boundary included: (i) elevation of the top of the stream surface = topographic elevation - 0.5 feet, (ii) streambed thickness = 0.5 feet (Wexler and Maus 1988), (iii) streambed conductivity = 27 feet/day ($1/10^{th}$ of horizontal hydraulic conductivity of 270 feet/day; Wexler and Maus 1988), (iv) starting head = 30 feet msl at the start-of-flow point, and (v) ending head = 0 feet msl at the mouth of the stream.

(a) (b)

Figure 3.21: Map showing (a) Swan River, and the (b) GHB simulating Swan River (in green)

3.2.4.7. Drains

All the streams located within the model domain – Beaverdam Creek, Little Neck Run, Yaphank Creek, and Carmans River – withdraw water from the groundwater system; 95% baseflow is groundwater (based on Peterson 1987). Therefore, the streams were simulated as drains (Figure 3.22). The elevation of the top of the stream surface was kept fixed at 0.5 feet below the elevation of the ground surface. The conductance of the streambeds was calculated as the product of the reach length of the drain in each grid cell and the conductance per unit length of drain in each grid cell. The DRN package in Visual MODLFOW was used to simulate the streams. The stream length was treated as a variable feature (discussed later).

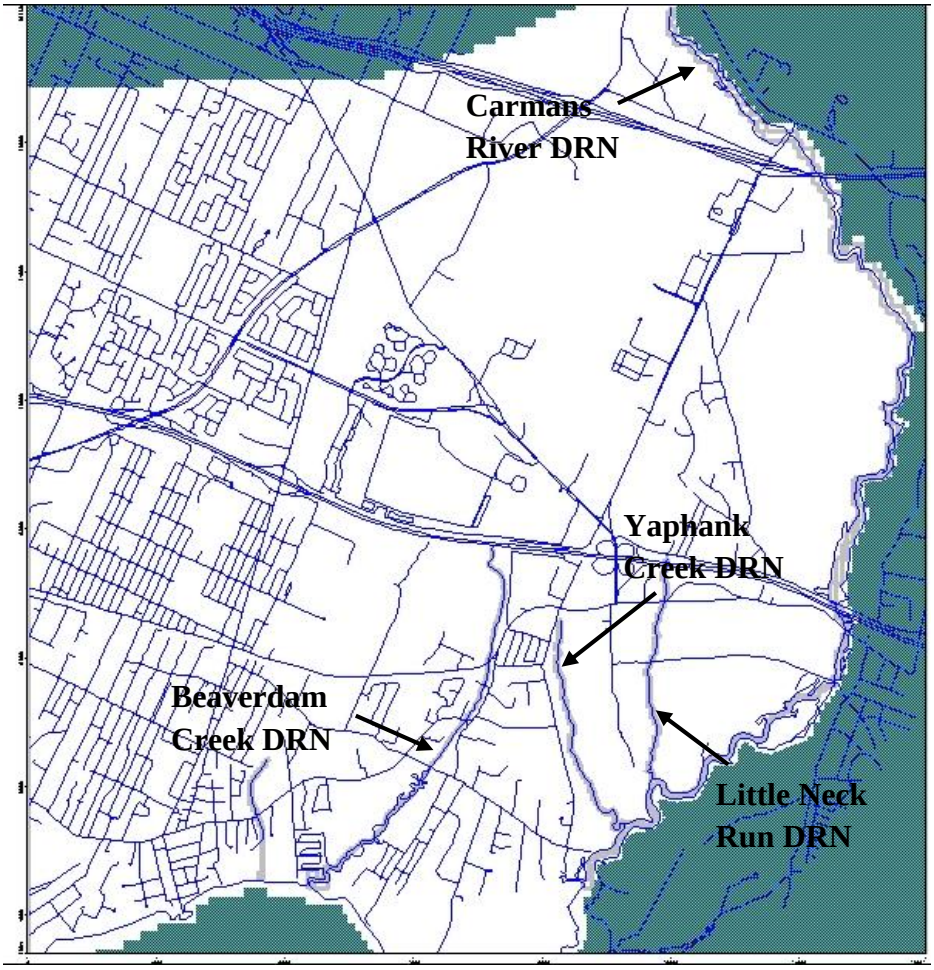


Figure 3.22: Drain features

3.2.4.8. Fixed Features

Fixed features are those model features assigned either a fixed value or a state.

- The PSU included in the model domain hydraulically separates the UGA from the underlying Magothy aquifer. The effectiveness of this separation depends on its water bearing properties, its extent, and its thickness. The extent and the water bearing properties were simulated as variable features (discussed later). A uniform thickness of 10 feet was assigned to the PSU in L4b.
- Hydraulic conductivity of the Magothy aquifer decreases from west to east. This pattern is concurrent with increases in thickness and content of fine material of the Magothy aquifer in the same direction (Gerathy and Miller 1985). McClymonds and Franke (1972) calculated the average horizontal hydraulic conductivity of the aquifer to be 54 feet/day with a range of 27 feet/day to 134 feet/day. Soren (1971) suggested a conductivity value of 67 feet/day for Nassau and Queens Counties, while Isbister (1962) suggested 268 feet/day for the same area. Franke and Cohen (1972) inferred a value of 50 feet/day. Here, the horizontal hydraulic conductivity (K_h) of the Magothy aquifer was fixed at 60 feet/day.
- The anisotropy ratio was fixed at 10:1 for the UGA (based on Lindner and Reilly 1983) and the PSU (Eckhardt and Wexler 1986). For Magothy aquifer the ratio was fixed at 30:1 based on Lindner and Reilly (1983).
- Direct measurements of the thickness of the streambed were not done or available. Here, the thickness of the streambed was kept fixed at 0.5 feet for the streams.
- The rate of precipitation was fixed at 48 inches/year. This value was approximated from the average annual precipitation rate of 48.3 inches/year for the period 1949-2013 for the monitoring station in Upton, NY.
- The rate of recharge was kept fixed at 24 inches/year: 50% of the average annual precipitation (48 inches/year) (based on Peterson 1987).
- The rate of evapotranspiration was kept fixed as 24 inches/year (based on Peterson 1987). The extinction depth was fixed at 3 feet below the ground surface, that is, the ground water evapotranspiration at the rate of 24 inches/year when the distance between the ground surface and the water table was less than 3 feet.

- A public water supply well – S33826 – located about 1,500 feet to the west of the western perimeter boundary of the landfill was simulated (Figure 3.23). The well is screened at -93 to -95 feet msl into the shallow Magothy aquifer. The pumping rate was kept constant at about 18,100 gallons per day or 12.5 gallons per minute, based on the pumping rate provided by Wexler (1988b).

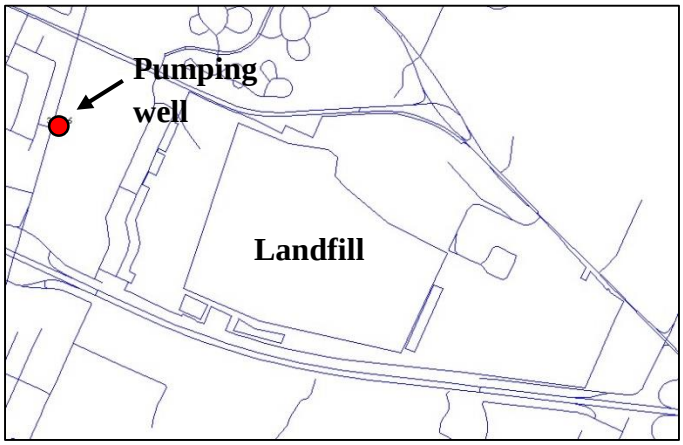


Figure 3.23: Pumping well

3.2.4.9. Other Fixed Features

In addition, the following features were represented by either a fixed state or value (Table 3.3).

Feature	Description	Value / State
Flow type	Density of fluid	Saturated (constant density)
Total porosity	Fraction of soil void of material	0.33 (Wexler 1988a)
Effective porosity	Porosity available for fluid flow	0.3 (Wexler 1988a)
Specific yield (Sy)	Storage or release of water from pore spaces due to change in water levels	Sand [27-21] (avg) (dimensionless) (Fetter pg. 79)
Specific storage (Ss)	Storage or release of water from pore spaces due to change in storage units	~0.0001 / ft (Fetter pg. 101)

Table 3.3: Other fixed features

3.2.5. Variable Features

Eight model features were considered as "variable features" in the model, that is, the uncertainty associated with these model features was recognized and incorporated into the modeling exercise. The uncertainty associated with variable features was represented by two or three select variations (states) of that feature. For example, some variables features were represented by a dichotomous "yes" – "no" variation. Other variable features' were represented by variable numerical values; for example, the horizontal hydraulic conductivity (K_x) for the UGA was represented by three different values: 300 feet/day, 250 feet/day, 200 feet/day. Each variable feature was assigned an alphanumeric code. For example, code "V12" indicated the variable feature 1 (V1): "top surface of the PSU", while the number "2" indicates second of the variable feature: "interpolated surface". Table 3.4 summarizes the 8 variable features and their states.

Code	Variable Feature	State 1	State 2	State 3
V1	Bottom of layer 1	Uniformly thick	Variably thick	N/A
V2	Bottom of layer 2	Uniformly thick	Variably thick	N/A
V3	Extent of the PSU	2-zone	3-zone	N/A
V4	Recharge (local)	Natural	Via Recharge Basins	No recharge
V5	Stream segmentation	Yes	No	N/A
V6	Kh – UGA (feet / day)	High	Medium	Low
	L1	300	250	200
	L2	250	200	150
	L3	200	150	100
V7	Top surface of the PSU	Uniform surface	Interpolated surface	N/A
V8*	CHD boundary at the northern edge	40'	42'	38'

Table 3.4: Variable features and their states (* variable feature representing the aleatory uncertainty in the model)

The variable features were divided into two groups on the basis of the nature of model uncertainty (epistemic or aleatory) (Figure 3.24). Seven variable features – V1, V2, V3, V4, V5,

V6, V6, V7 – represented the epistemic uncertainty in the model. Variable feature "V8" represented the aleatory uncertainty associated with the groundwater system (fluctuations in the groundwater levels in the study area).

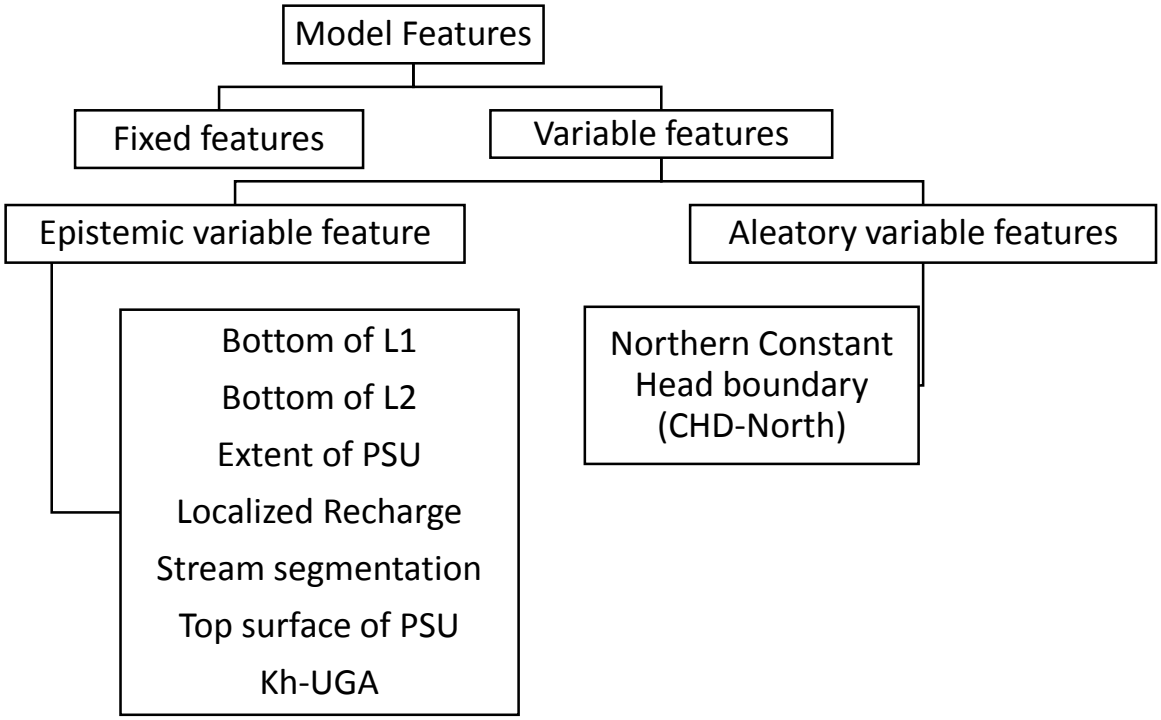


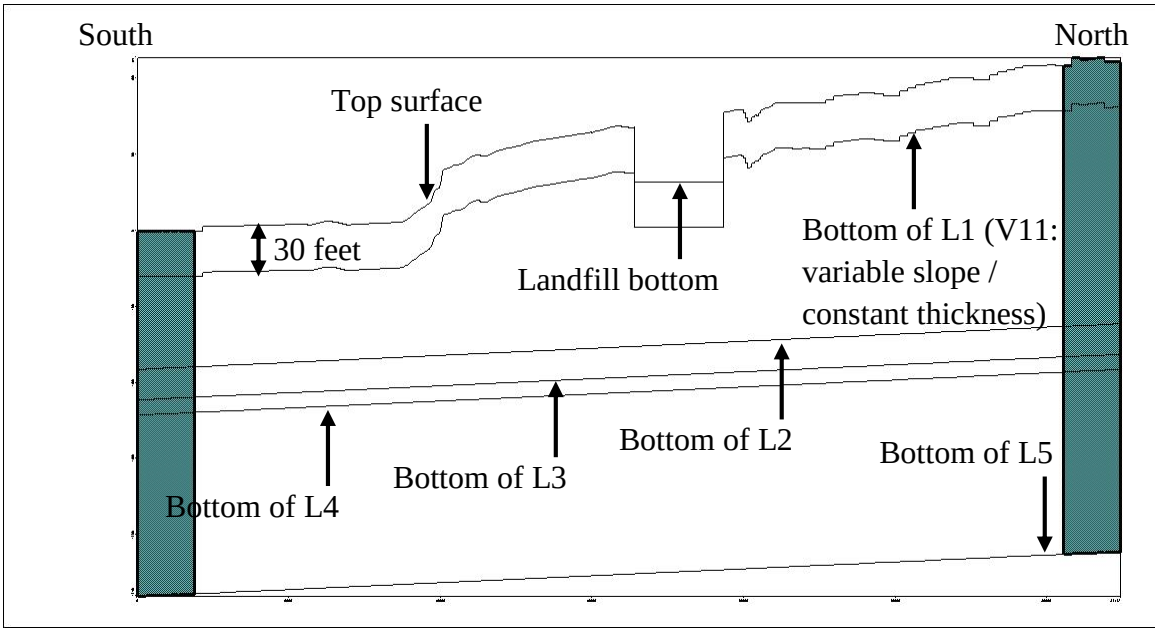
Figure 3.24: Classification of model features

3.2.5.1. V1: Bottom of Layer 1 (L1)

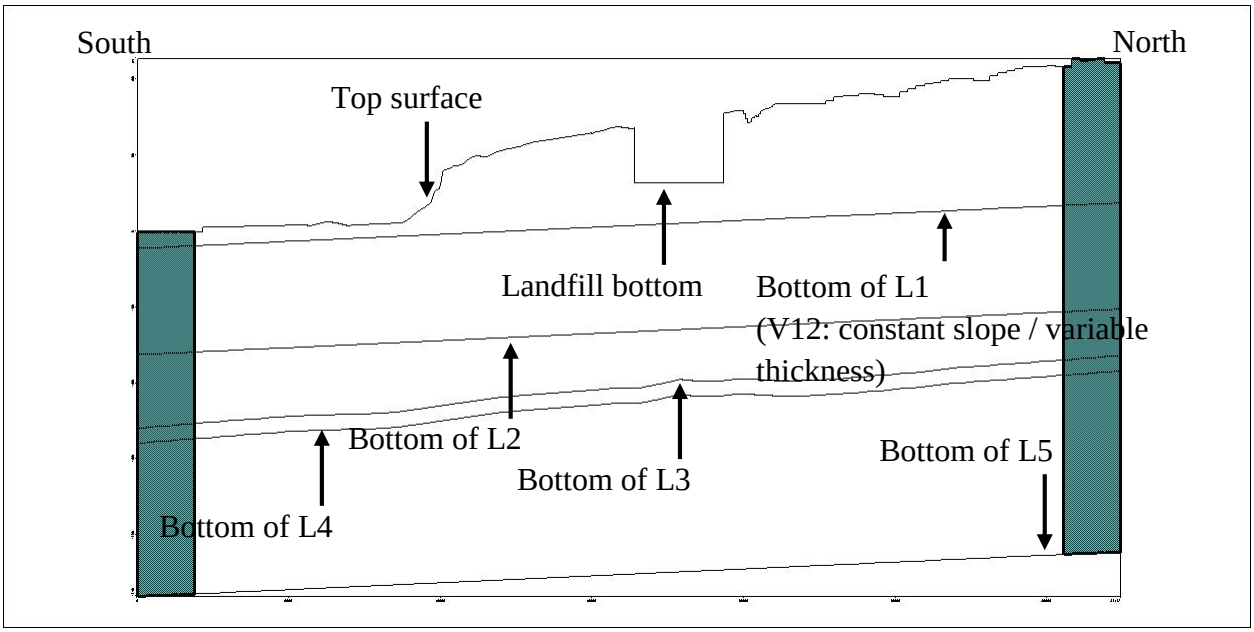
The sediments in the UGA show a downward fining trend (Dvirka and Bartilucci 1994a; Lindner and Reilly 1983; McClymonds and Franke 1972); however, the demarcations among these sediment zones are uncertain. The uncertainty in the positioning of the bottom of L1 represents the boundary between the highest permeability zone and the intermediate permeability zone. Changing the position of the layer bottom alters the thicknesses of layer 1 and that, in turn, changes the transmissivity of the layer (the product of layer thickness and its hydraulic conductivity). Altering the layer bottom also changes the layering for well screens which can affect the heads in the wells. This uncertainty was represented by two alternative states: V11 and V12 for the first layer (L1).

V11: L1 was assigned a constant thickness of 30 feet. Consequently, the bottom of L1 mirrored the topographic profile (Figure 3.25-a). A value of 30 feet was chosen by trial-and-error such that it prevented model cells from becoming dry.

V12: The bottom of L1 was simulated as a constant slope with a northeasterly strike and southeasterly tilt at a gradient of 0.067 feet/feet from the beginning elevation of 25 feet (Figure 3.25-b). The constant slope configuration was consistent with the constant slope of the UGA depicted in the generalized depiction of Long Island's geologic profile mentioned in Figure 3.5 above.



(a)



(b)

Figure 3.25: Profile view of a typical model domain along column 130 showing (a) constant thickness / variable slope (V11 state), and (b) variable thickness/constant slope (V12 state) of L1

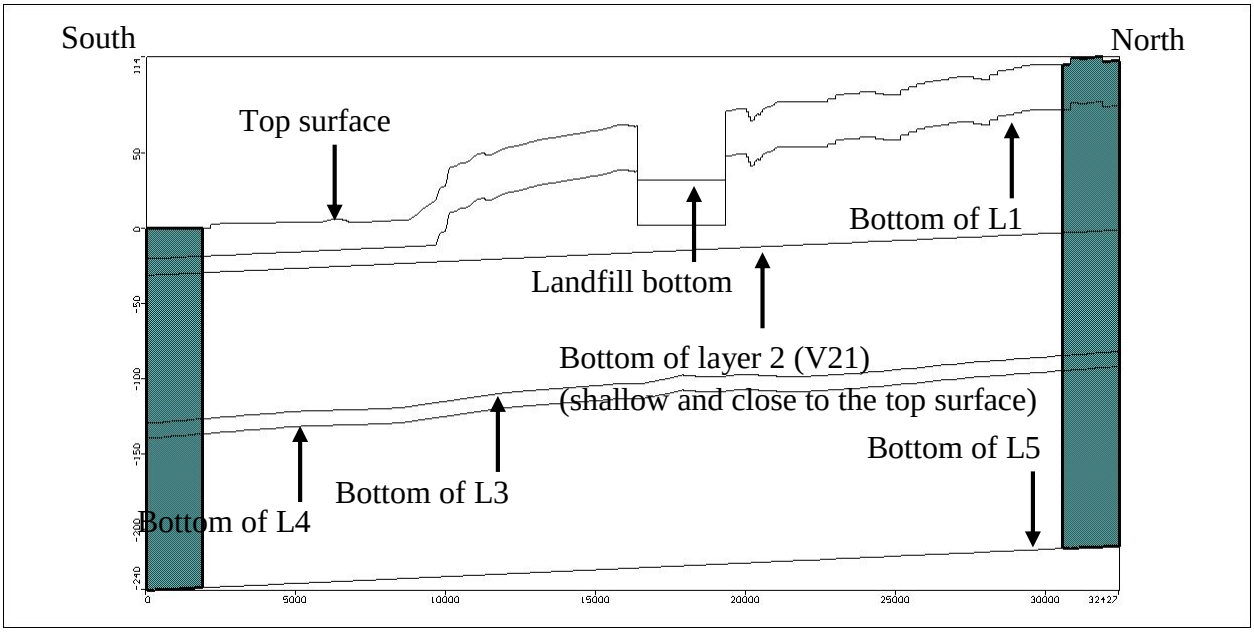
3.2.5.2. V2: Bottom of Layer 2 (L2)

Similarly the bottom of layer 1, the demarcations between the intermediate permeability zone and the low permeability zone of the UGA, are uncertain. Therefore, the bottom of layer 2 was also represented by two states: V21 and V22.

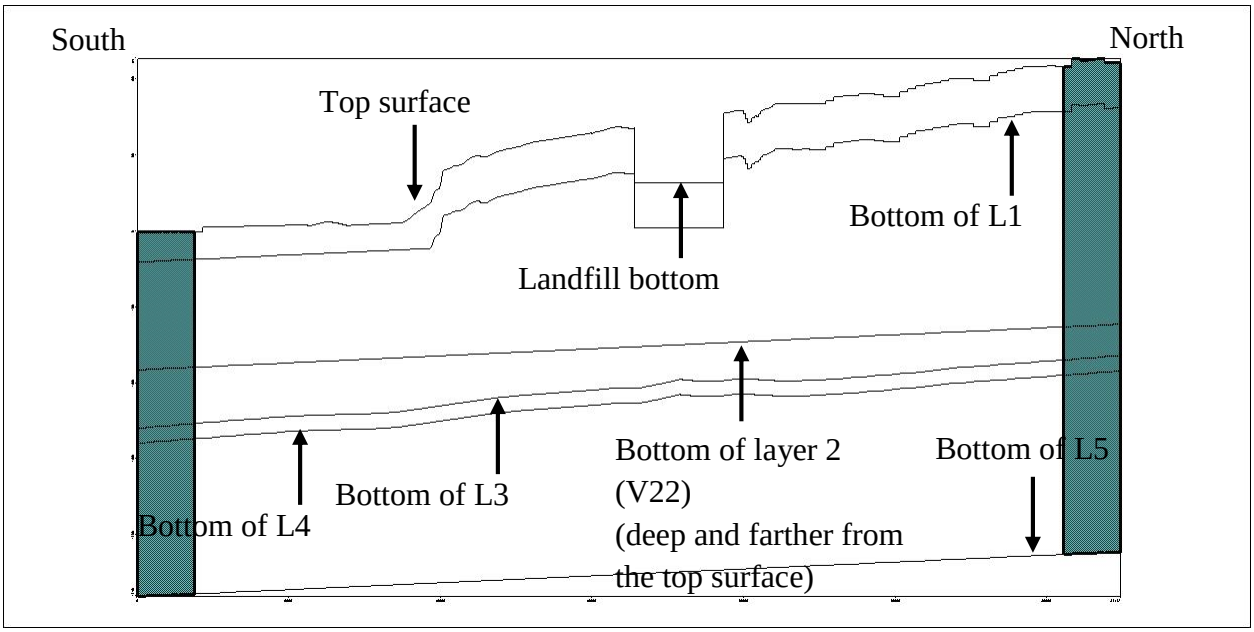
V21: The bottom of L2 was depicted as a constant slope with a northeasterly strike and a southeasterly tilt at a gradient of 0.067 feet/day from the initial elevation of -25 feet msl at its northern starting point. As a result, the bottom of L2 appeared shallow and closer to the top surface (Figure 3.26-a).

V22: The bottom of L2 was depicted as a constant slope with a northeasterly strike and a southeasterly tilt at a gradient of 0.067 feet/day from the initial elevation of -75 feet msl at its northern starting point. As a result, the bottom of L2 appeared to be deeper and farther from the top surface (Figure 3.26-b).

Similar to the alternate positionings of bottom of L1, changes in the position of bottom of L2 increases (or decreases) the thicknesses of L2 and L3. This changes he transmissivity of the aquifer and the screen zones of the wells. These changes could affect the heads.



(a)



(b)

Figure 3.26: Profile view of the model domain along column 130 showing bottom of V2 (a) closer to L1 (V21 state), and (b) closer to L3 (V22 state)

3.2.5.3. V3: Extent of the PSU

The extent of the PSU north of the landfill site is uncertain (Aphale and Tonjes 2013). This uncertainty was represented by two states: "V31" and "V32".

V31: This state simulated the absence of the PSU to the north and underneath the landfill. Here, the hydraulic conductivity of Zone 2 was changed to that of the bottom of layer 3 (L3) of the UGA. This is equivalent of the PSU being absent at Zone 2. The UGA is directly hydraulically connected with the Magothy aquifer in this state at zone 2. The PSU begins south of the landfill 825 at Zone 3. In this state, the hydraulic conductivity of Zone 3 was set at 0.1 feet/day (based on Wexler and Maus 1988).

V32: The PSU begins at Zone 2 and continues to Zone 3. The hydraulic conductivity progressively reduces from Zone 2 and Zone 3; the K_h of Zone 2 was set at 0.1 feet/day, while the K_h of Zone 3 set at 0.01 feet/day. The anisotropy ratio was set at 1:10 and the K_z of both zones was changed accordingly (Table 3.5).

State	Zone	K_x (feet / day)	K_y	K_z
V31	Zone 1	200 / 150 / 100	200 / 150 / 100	20 / 15 / 10
	Zone 2	200 / 150 / 100	200 / 150 / 100	20 / 15 / 10
	Zone 3	0.1	0.1	0.1
V32	Zone 1	200 / 150 / 100	200 / 150 / 100	20 / 15 / 10
	Zone 2	0.1	0.1	0.1
	Zone 3	0.01	0.01	0.001

Table 3.5: Two states of the variable feature V3

3.2.5.4. V4: Local Recharge

Local recharge represented the influx of precipitation into the ground at the landfill. The bottom of the landfill mounds is lined by artificial liners to prevent recharge from percolating out of the landfill mounds. The exact status and efficacy of the liner is unknown. It is suggested that Cell 1 and Cell 2 of the landfill are leaky sections of the landfill given the detection of contaminated groundwater in the 1980s (Wexler 1988a). The uncertainty in the liner-recharge relationship was represented by three states of variable feature V4: V41, V42, and V43.

V41: Here, it was assumed that the landfill liners are non-existent or ineffective and the rate of recharge at the landfill mounds occurred as that in the pre-development conditions at the rate 24 inches/year.

V42: In this state it was assumed that the recharge occurred through on-site recharge basins (Figure 3.27). This would occur if landfill caps (installed in 1993 for Cells 1-3) prevented precipitation from entering the landfill, but instead caused runoff down the slopes into the recharge basins. On-site recharge basins cause local re-distribution of recharge that, in turn, may affect local groundwater levels particularly in the shallow wells at the southern-eastern periphery of the landfill. The precipitation was redistributed to the three recharge basins on the basis of the fractional share of their individual recharge rates (Table 3.6).

Recharge Basin	A	B	C	Total
Area (ft^2)	88,849	103,868	90,553	283,270
Volume (ft^3/year)	5,102,888	5,925,935	5,267,498	16,296,321
Fraction of total volume	0.31	0.36	0.32	1.00
Recharge rate (ft/year)	57.43	57.05	58.17	--
Recharge rate (inches /year)	689.16	684.6	698.04	--

Table 3.6: Fractional recharge rate of recharge basins

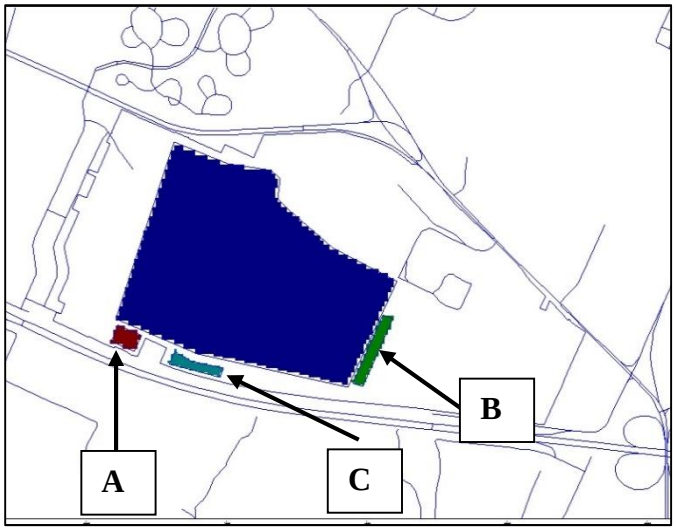


Figure 3.27: The recharge basin (a) A (green) (b) B (brown), and (c) C (blue-green) and the landfill mound (blue).

V43: This state simulates the conservative scenario where it was assumed that the liner system is fully functional (as opposed to being non-existent as in V1) and completely prevents any recharge to the ground (RCH = 0). The leachate generated is collected into the storage tanks through a network of leachate collection pipe, and then is sent off-site for further treatment.

Figure 3.28 depicts three recharge scenarios or the states of variable feature V4.

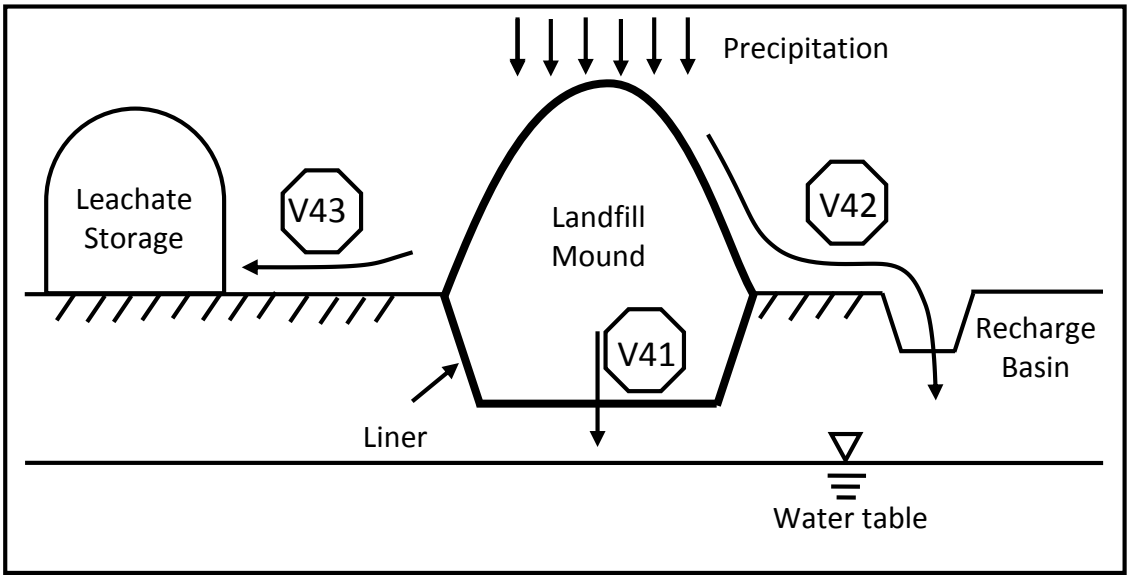


Figure 3.28: Three states of variable feature V4

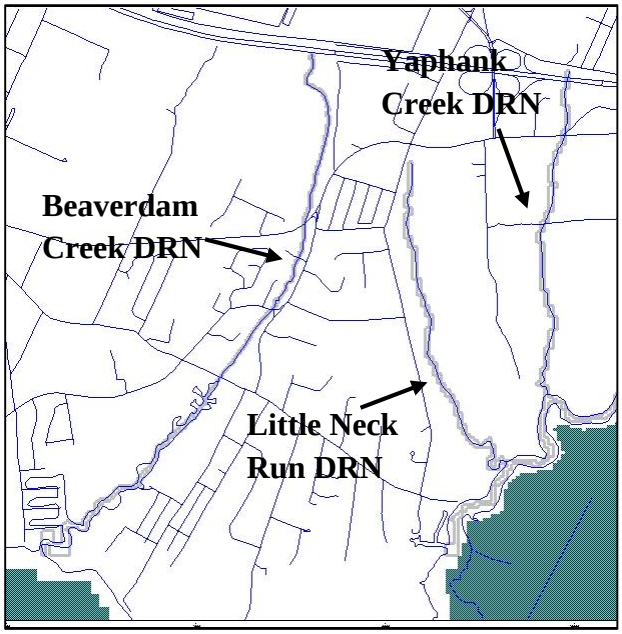
3.2.5.5. V5: Stream Segmentation

Two states were simulated with regard to the stream settings.

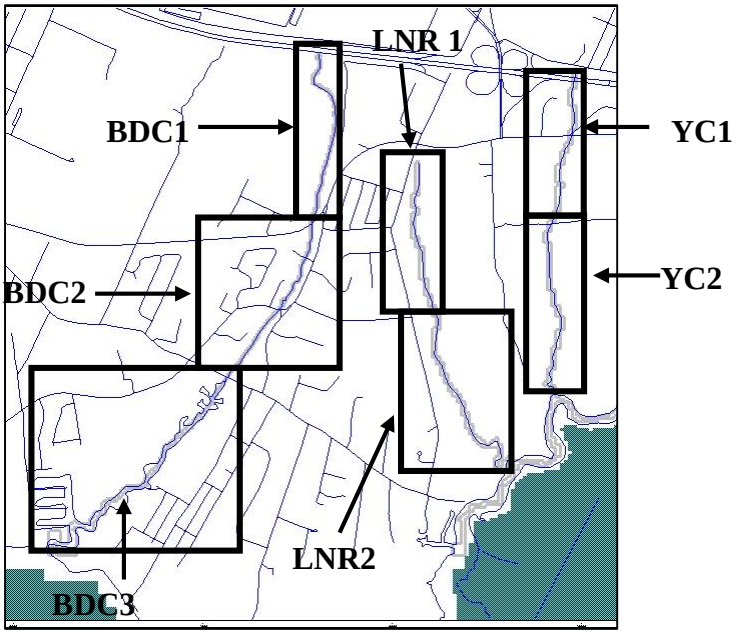
V51: The drains were simulated as a single polyline features whose width broadened from the headwaters to the mouth of the stream via linear interpolation (from, 2 feet at the start-of-flow to 30 feet for the Beaverdam Creek, for example) (Figure 3.29-a-b).

V52: The steams were divided into multiple segments that were then simulated as an individual drains (Figure 3.29-c-d). For example, Beaverdam Creek was divided into three segments; the first segment simulated the intermittent headwaters of the Creek (BDC1), the second segment simulated the perennial stream segment (BDC2), and the third segment simulated the tidal sections of the Creek (BDC3) (Figure 3.29-c). The width of these segments broadened from the point start-of-flow to the end point of each segment; from 2 to 5 feet (BDC1), 5 to 10 feet (BDC2), and 10 to 30 feet (BDC3). Similarly, Yaphank Creek and Little Neck Run were divided into two segments (Figure 3.29-c). The Carmans River was divided into five segments (Figure 3.29-d).

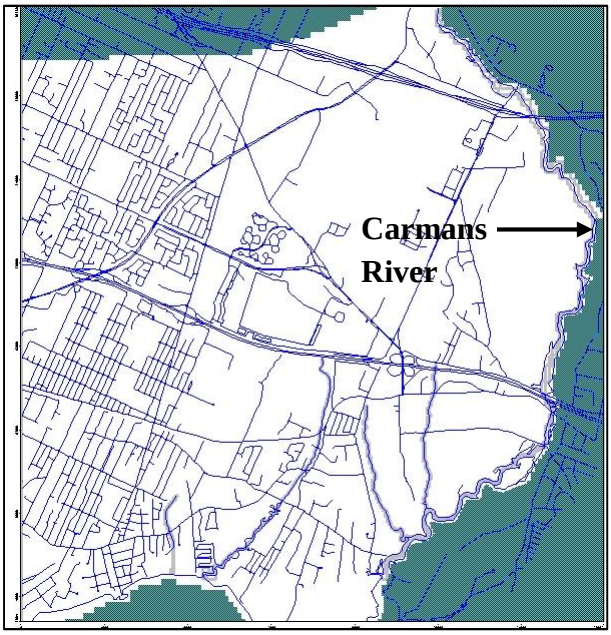
Linear interpolation is less resource intensive because it requires data at the start-of-flow and at the mouth of the river (as opposed to at the starting and at the end point of each segment, as in case of state V52).On the other hand, stream segmentation allows discretization of the streams into multiple segments so that the characteristics of the streams, such as their dimensions, streambed thickness, and conductance, can be individually and better set for each segment.



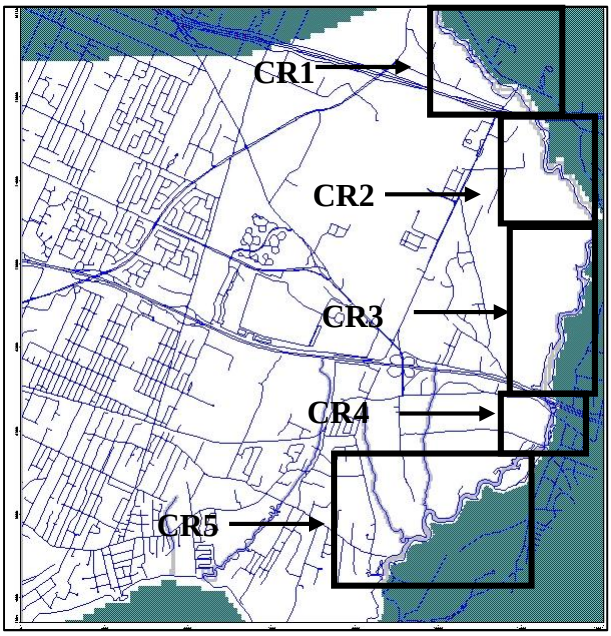
(a)



(c)



(b)



(d)

Figure 3.29: Drain (DRN) features in the model showing the (a) V51 state for Beaverdam Creek, Little Neck Run, and Yaphank Creek, (b) V51 state for Carmans River, (c) V52 state for Beaverdam Creek, Little Neck Run, and Yaphank Creek, and (d) V52 state for Carmans River

3.2.5.6. V6: Hydraulic Conductivity of UGA

The deposits in the upper sections of the UGA are generally coarse and readily yield water, while the deeper sections of the UGA have better sorted sands with low K (downward fining). The exact values of the conductivities are uncertain. Therefore, three sets of K values – High K, Medium K, and Low K – were used to represent the downward fining of the UGA (Table 3.7). The anisotropy ratio was fixed at 1:10.

Layer	High K (V61)	Medium K (V62)	Low K (V63)
L1	300	250	200
L2	250	200	150
L3	200	150	100

Table 3.7: Sets of K values for the three layers of the UGA (in feet/day)

As mentioned in section 3.1.1.3, the values for the hydraulic conductivity vary in the UGA with location, with depth, and with the type of technique used to calculate the conductivity values. Several combinations of conductivity values could be developed and used in the model. However, the range of conductivity values represented by the three conductivity sets in Table 3.7 above represent a conservative range for the conductivity at each UGA layer.

3.2.5.7. V7: Topography of the PSU Surface

Just as the horizontal extent of the PSU is uncertain, the surficial profile of the PSU surface is uncertain as well. Therefore, the variations in the topography of the PSU was represented by two states: as a constant southeasterly slope (V71), or, as an undulating surface interpolated from various geologic boring logs (V72).

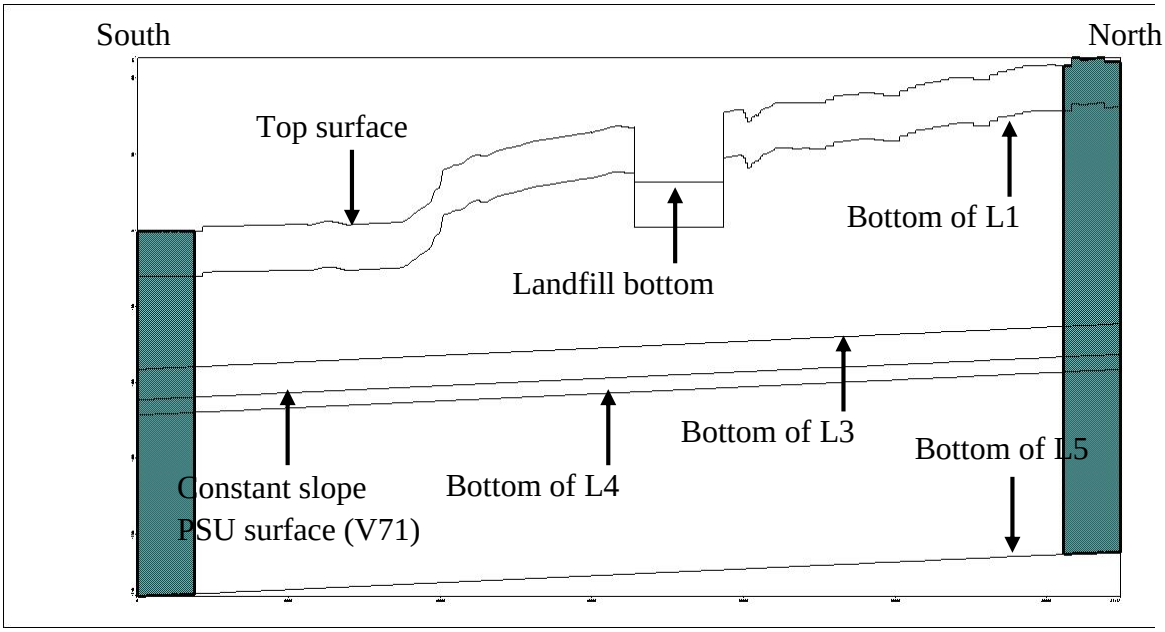
V71: The top surface of the PSU was depicted as a constant slope dipping in southeasterly direction, consistent with the regional geologic descriptions (Figure 3.30-a). A "dip and tilt" feature in Visual MODFLOW 4.2 was used to depict the surface. The top surface depicted as a constant slope with a northeasterly strike and southeasterly tilt at a gradient of 0.067 feet/day from the initial elevation of -90 feet msl at its northern starting point.

V72: The top surface of the PSU was interpolated based on the lithological descriptions in the boring logs obtained at 29 locations in the study area (Figure 3.30-b) (Table 3.8). A boundary or "cut-off" between the bottom of the UGA and the top of the PSU was determined from these boring logs. The interpolation was executed using the natural neighbor method using the built-in surface interpolation feature of the Visual MODFLOW v. 4.2. The lower surface of the PSU layer had an identical topographic profile as that of the upper surface because the thickness of the PSU was assumed to be 10 feet.

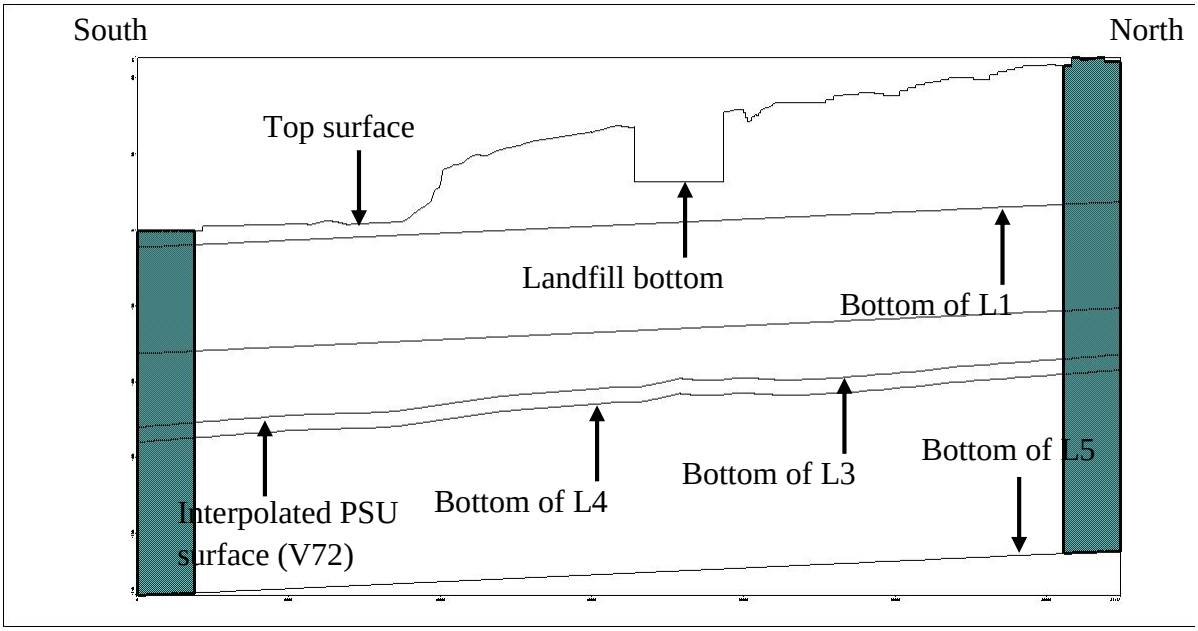
The constant slope depiction of the top surface of the PSU was consistent with the generalized geologic profile of Long Island mentioned in Figure 3.5 above. On the other hand, the interpolated surface of the PSU in state V72 is based on the lithological descriptions in the boring logs obtained at 29 locations in the study area. Although location-specific evidence is preferable, the distribution of these borings across the model domain (as well as with respect to the depth) was limited. In addition, the interpolation methods calculated surface elevations for the area bigger than the area enclosed by the outermost set of boring locations. This extrapolation may change if additional set of boring locations are included in the calculation. Hence, two states were developed to depict the top surface of the PSU.

Boring	Well #	X (feet)	Y (feet)	Cut-off (feet msl)
1	47438	1288146	253593.6	-75
2	65905	1348995	244412.8	-100
3	49018	1275437	230462.3	-175
4	71882	1302356	236569.8	-130
5	62022	1261698	227872.4	-110
6	28208	1245315	213419.9	-75
7	47035	1243301	221798.8	-100
8	66184	1246772	228111.1	-130
9	52493	1297866	221025.8	-100
10	29492	1244955	239730.5	-80
11	69364	1278334	219870.3	-110
12	52944	1274799	238754.1	-90
13	46713	1313550	233379.1	-105
14	47024	1338531	223913.5	-80
15	9349	1260358	190105.9	-150
16	18846	1242637	182927.5	-175
17	129174	1287234	223976.7	-125
18	B18	1279123	231813.7	-100
19	B20	1279927	229403.6	-102
20	B21	1278396	228811.2	-95
21	72813M	1280972	229768.6	-115
22	72814M	1282707	225829.3	-110
23	PB24	1281663	230041.8	-85
24	11M	1278880	229157.3	-95
25	10D	1279395	229555.8	-105
26	5D	1277416	231833.8	-100
27	12D	1279022	232538	-95
28	4D	1279898	229192.7	-100
29	8D	1279224	230571.4	-95

Table 3.8: Boring locations and the cut-off points



(a)



(b)

Figure 3.30: Interpolated slope of the PSU profile view at column 130 showing (a) uniform slope (V71 state), and (n) interpolated slope (V72 state)

3.2.5.8. V8: Northern Constant Head Boundary (CHD-North)

The CHD boundary, along with precipitation, represented the source of water in the model domain. Changes in either would change the water input to the model domain that would then change the groundwater levels. The precipitation value was kept fixed (RCH = 48 inches/year). Therefore, the historical fluctuations in water levels were simulated by changing the water influx in the model domain through the northern constant head boundary (CHD-North).

This is the only variable feature in the model that represented the aleatory uncertainty associated with the groundwater systems: fluctuations in the groundwater levels. The variations in the groundwater levels were simulated by altering the value of the constant head (CHD) boundary at the northern edge of the model (CHD1) to represent the high, median, and low groundwater levels. The values were based on the historic annual water table contour maps of Long Island for the period 1983-2010 (Monti et al. 2013; Monti and Busciolano 2009; Busciolano 2002; Busciolano et al. 1998; Doriski 1987).

Three states (V81, V82, and V83) were simulated:

V81: CHD1 = 42 feet representing high groundwater levels,

V82: CHD2 = 40 feet representing median groundwater levels, and

V83: CHD3= 38 feet representing low groundwater levels.

3.3. The Area Metric

One approach to access the failure of deterministic model comparisons is to express the value of both the observed as well as the simulated quantity as distribution functions. In this way, the uncertainty in these quantities can be explicitly acknowledged, an uncertainty that may be either aleatory and / or epistemic in nature (Ferson et al. 2008). This is generally accomplished through the area metric.

3.3.1. Empirical Cumulative Distribution Function (ECDF)

If a variable, X, such as groundwater head, assumes different values x_i (i=1,…, n), then each individual value can be associated with a probability of its occurrence. These individual values can be collated in a probability distribution function (PDF) that can be alternatively depicted as a cumulative distribution function (CDF). The CDF represents a cumulative probability that X will be less than or equal to each possible sampled value from a population of values of *x*: F(*x*) = P[X≤*x*] (Morgan and Henrion 2006, p. 74). A function $\hat{F}(x)$ is a CDF if and only if (i) $\lim_{x\to-\infty} \hat{F}(x) = 0$ and $\lim_{x\to\infty} \hat{F}(x) = 1$, (ii) $\hat{F}(x)$ is a non-decreasing function of *x* (that is it is a monotonic function with a non-negative slope), (iii) $\hat{F}(x)$ is right continuous, and (iv)

$$\hat{F}(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 0 \end{cases}$$

The CDF can either have infinitely long tails or it can be truncated to finite interval ranges of *x*s. Typically, a CDF is constituted by infinitely large number of values that are randomly selected from an interval range. In reality, however, the number of groundwater head observations made are not infinite. Therefore, those observation data points can be arranged in an increasing order in a monotonically increasing step function called the empirical cumulative distribution function (ECDF). It is a CDF constituted using observed empirical data. The ECDF is constituted using a limited number of empirically observed quantities. It is a non-parametric graphical representation of the probability distribution. No prior assumption is made about the form of a PDF, and no parameter (e.g. mean, standard deviation) is selected to specify such a distribution. If the number of observations are very large to infinite in number, then the ECDF may estimate the true CDF for a variable. The data points are ordered in an increasing order

(from the smallest to the largest) on the vertical axis (the probability axis). Each point represents a step positioned on the horizontal axis at the *x* values. Thus, an ECDF appears as a monotonically increasing discrete distribution divided into *n* vertical steps of equal length from 0 to 1 (Ferson et al. 2008). For example, Figures 3.31-a shows the head observations recorded at well S3529 between 1976 and 2010. Figure 3.31-b shows the PDF of the 446 observations and Figure 3.31-c shows the corresponding ECDF consisting of 446 vertical steps of equal length.

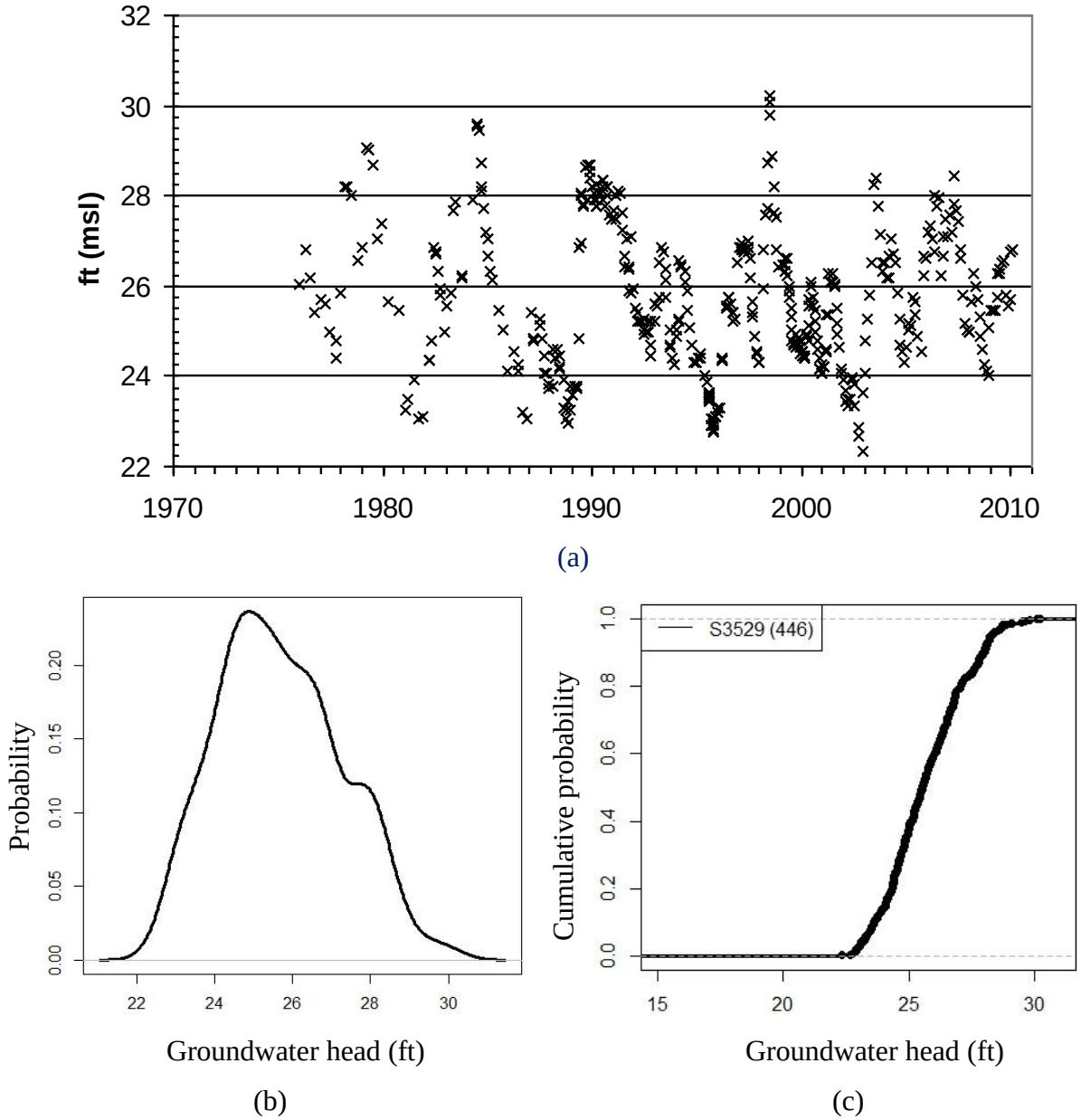


Figure 3.31: (a) Water table heights at well S3529 (1975-2010) (in ft. msl), and (b) the corresponding PDF, and (c) corresponding ECDF

The ECDF is a monotonically increasing function. Therefore, the inherent order in the data, such as a chronological ordering in case of head observations taken over a period of time, is overridden by an increasing order of magnitude; however, the central tendency and the dispersion of these data is retained (Ferson et al. 2008). In the figure above, the largest value of the head (30.20 feet, June 1998) occupied the top-right position in the ECDF, while the smallest value of the head (22.32 feet, November 2002) occupies the bottom-left position in the ECDF, overriding their chronological arrangement.

The ECDF can be constructed with as little as a single datum, in which case the ECDF takes the form of a "spike" distribution at the datum value. In this case, all the values less than the datum have a probability of zero, while all the values greater than the datum will have a probability of one (Figure 3.32-a). The steps in an ECDF increase with the increase in the data points. For example, Figure 3.32-b shows ECDF with 10 data points, while Figure 3.32-c shows ECDF with 5 data points. The ECDF smoothens into a curve with sufficiently large number of data points (Figure 3.32-d).

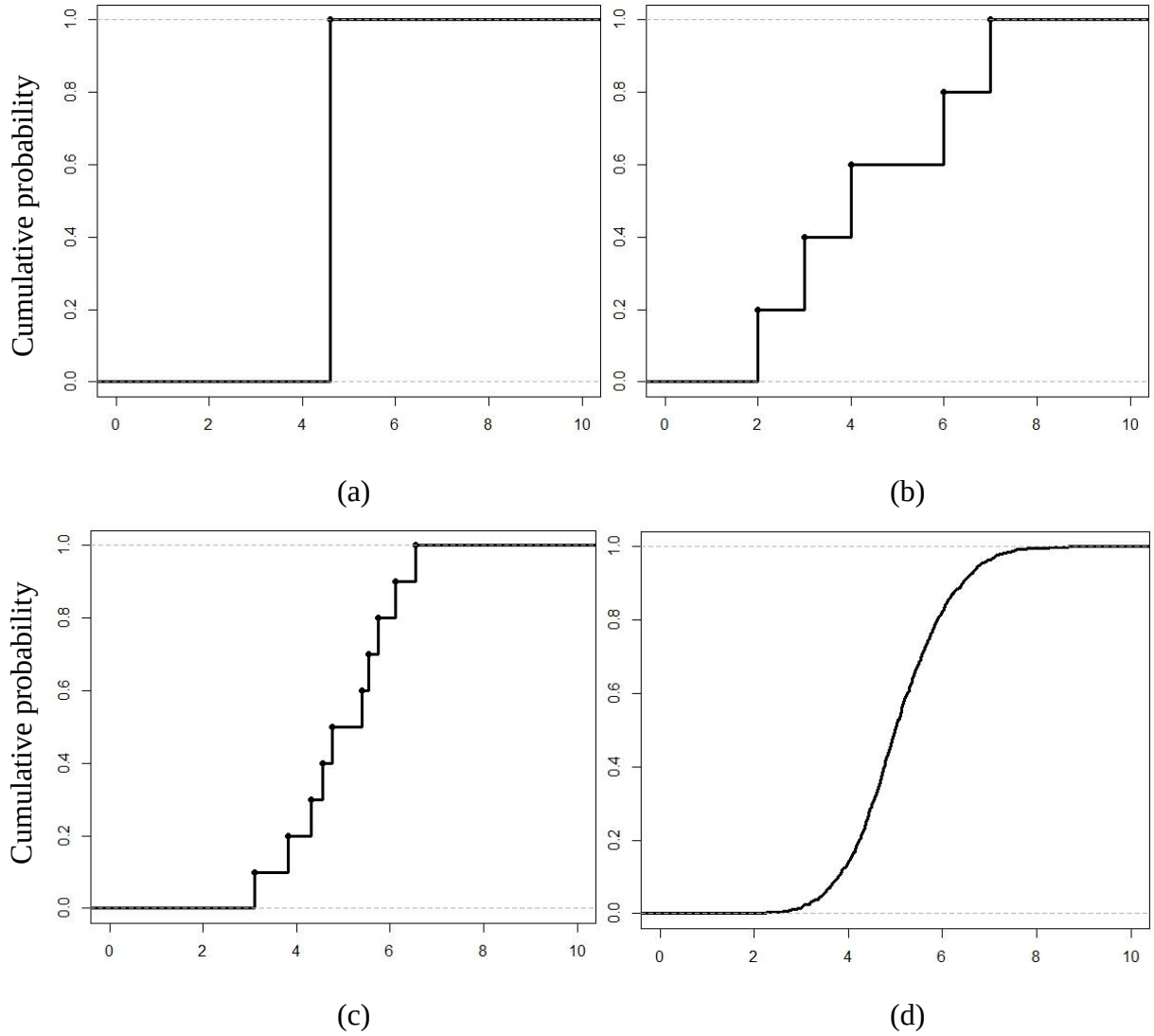


Figure 3.32: ECDFs where (a) n=1, (b) n=5, (c) n=10, and (d) n → ∞

3.3.2. The Area Metric

The value of the area metric is the integral of the absolute value of the difference between the ECDF of the observed data and the ECDF of the simulated data (Roy and Oberkampf 2011; Ferson et al. 2008) (Equation 3.4).

$$d\,(F,Sn) = \int_{-\infty}^{\infty} |\hat{F}(x) - Sn(x)|dx \qquad (3.4)$$

where,

$\hat{F}(x)$ = the ECDF of the observed values ($ECDF_{observed}$)

$Sn(x)$ = the ECDF of the simulated values ($ECDF_{simulated}$)

The area metric indicates that the level of disagreement between the observed values and the model-simulated values expressed as their ECDFs. The area metric is analogous to the Euclidean distance (*d*); as with *d*, smaller area metric values suggest lesser disagreement between the observed and the simulated data (Ferson et al. 2008). Models with smaller area metric values can be considered to have better (replicative) validity with respect to the observed data.

Consider one observed and simulated datum depicted as their ECDFs (Figure 3.33). The ECDFs appear as a "spike": a CDF with zero variance. The area of the rectangular area enclosed between these ECDFs is described as the area metric (*A*) (Ferson et al. 2008).

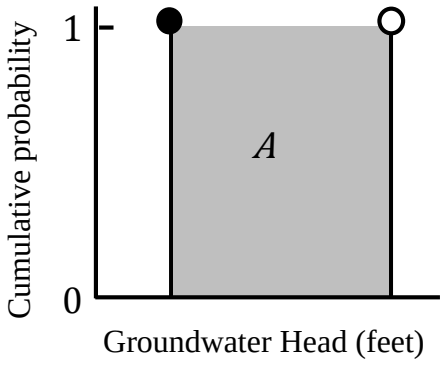


Figure 3.33: $ECDF_{observed}$ (solid circle) and the $ECDF_{simulated}$ (open circle); A = Area metric (feet)

If instead of a single observation a large number of observations are made at the observation location, and an equally large number of simulated model outputs are generated, then their ECDFs appear as smooth curves (Figure 3.34). The dispersion or spread with in these

ECDFs depends on the variations in their constituent data. The dispersion within the observed data ECDF is a result of the inherent variability in the groundwater system as reflected by fluctuating groundwater heads. This variability represents the aleatory uncertainty associate with the groundwater system.

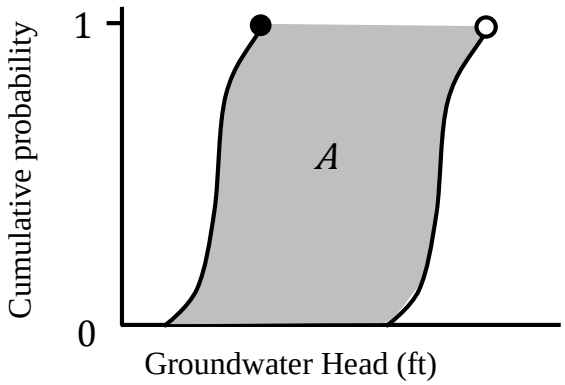


Figure 3.34: Comparison between the $ECDF_{observed}$ (solid circle) and the $ECDF_{simualted}$ (open circle); A = Area metric (feet)

The uncertainty in the model outputs can be represented by CDFs of model simulated values (Roy and Oberkampf 2011). The dispersion in the simulated data ECDF arises from range of outputs generated by the model due to the changes the model configuration in response to the corresponding changes in states of the groundwater system. For example, if the groundwater levels fluctuate from high to median to low levels, then the model configuration can be adjusted so as to produce three sets of simulated outputs, one corresponding to each state of the groundwater levels. These triplicate simulated outputs can then be collated into a model simulated ECDF. The simulated values are likely to vary with varying model configurations and this variation will reflect in the spread of the simulated data ECDF as shown in Figure 3.34.

In modeling studies, the count of observed data lies somewhere between a singular observation and an infinitely large observational data set. For instance, if the number of observation data points are three, then the ECDF of the observed data looks like a discrete step function (Figure 3.35). This discrete step function is considered to be reasonable estimate of the otherwise smooth curves that are generated by large number of simulation runs (Ferson et al. 2008). If equivalent number of model simulated values are generated, then these simulated

values can be collated into the model simulated ECDF. The area metric would be the area enclosed between these discrete step functions.

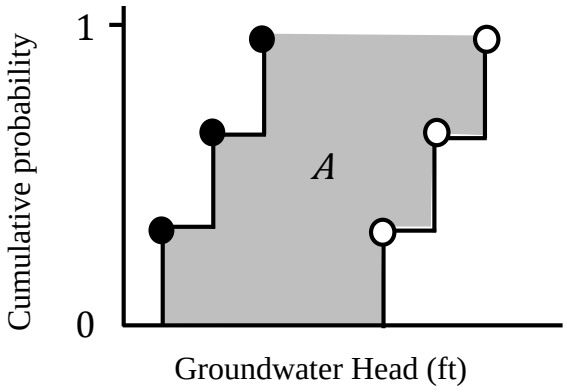


Figure 3.35: Comparison between the $ECDF_{observed}$ (solid circle) and the $ECDF_{simulated}$ (open circle) when n =3; *A* = Area metric (feet)

The above comparisons involve only one model. However, when multiple models are generated then each model could be represented by its own distinct ECDF. For instance, Figure 3.36 shows three model ECDFs. These ECDFs may occupy different positions on the horizontal axis, suggesting that the area between each one of these and the distribution of the observed data is different. Each comparison yields an area metric, these area metric values can be used to rank the models in terms of replicative validity (Ferson et al. 2008; Oberkampf et al. 2002). The position of model ECDFs over the horizontal axis represents the epistemic uncertainty, while the dispersion within any given ECDF is represents the aleatory uncertainty embedded within the ECDF.

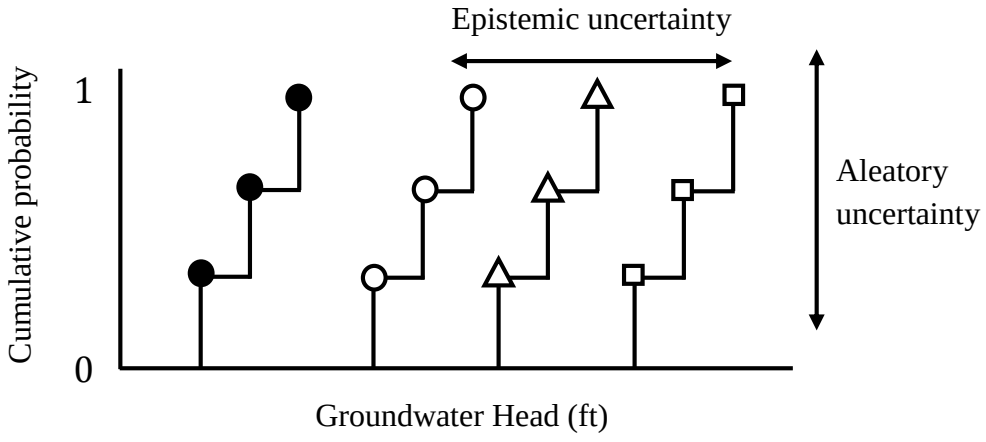


Figure 3.36: $ECDF_{observed}$ (solid circle) and the $ECDF_{simulated}$ from a model M1 (open circle), model M2 (open triangle), and model M3 (hollow square)

The dimensions of a geometric area are L^2. However, the height of this CDF is dimensionless because it is plotted on a probability scale. Therefore, the value of the area "A" enclosed by the two CDFs has a single length unit (L^1). The same as that of the observed data. Therefore, the area metric has in the same units as the observed data (if the simulated and the observed values are in feet then the units of the area metric are in feet as well) (Ferson et al. 2008; Oberkampf and Barone 2006).

The area metric satisfies the conditions of non-negativity, symmetry, triangle inequality, and identity of indiscernible for a distance function on area metric space (Ferson et al. 2008). The area metric is non-negative because it is an absolute measure of difference between the observed ECDF and the simulated ECDF. The absolute nature of the area metric also makes it symmetric because its value remains non-negative ($0 \leq A$). The triangular inequality indicates that $A_{M1} + A_{M1M2} > A_{M2}$ where A_{M1}, and A_{M2} are the area metric values calculated between the observed data the ECDF of model M_1, and the ECDF of M_2 respectively, while A_{M1M2} is the value the area metric calculated between the distribution of model M_1 and the distribution of model M2. The condition of identity of indiscernible means that the area metric can be zero if and only if the observed data ECDF and the simulated data ECDF are indiscernible from each other in terms of both the location and shapes (Ferson and Oberkampf 2009). The value of the area metric is generated by the function of the shapes of the distributional curves and not as the constituent variables of these distributions (Ferson et al. 2008).

As the aleatory uncertainty within the observed ECDF and the simulated ECDF approaches zero, the area metric reduces to simple difference (*d*) between the two ECDFs. This can also happen when the observed data and the simulated data are single point values represented as spike distributions, as shown in Figure 3.37 above (Ferson et al. 2008). However, if either one of the distributions is a curve, then the ECDFs will be discernible from each other and can never perfectly match ($A > 0$). Also, the area metric is unbounded: the area metric can take infinitesimally large values depending on the distance between the observed CDF and the simulated CDF (Ferson et al. 2008).

The concept of the validation area metric and its derivative, the probability bounds analysis (PBA), are applied primarily in risk assessment (Ferson 2001). Ferson and Tucker (2003) suggested that the PBA approach is suitable in environmental risk assessment problems and in assessment of risk of "high-consequence systems" where the risk associated with the

occurrence of the extreme events (worst or best-case scenarios) is to be estimated, such as safety at nuclear power reactors or at a the sub-surface nuclear waste repository (Oberkampf and Barone 2006). Rozell and Reaven (2011) assessed the likelihood of water contamination from natural gas extraction in the Marcellus Shale using probability bounds analysis (PBA). Ferson and Tucker (2006) demonstrated the application of the PBA in conducting sensitivity analysis when the spread of the probabilistic interval of a p-box can be reduced or "pinched" when better estimates (real values or precise probability distributions) are available. The current use of the area metric is unique to the validation for groundwater flow simulation models.

3.3.3. Application of the Area Metric

The generation of ECDFs and the calculation of the values of the area metric followed in three steps (Figure 3.37):

Step I: Compilation of ECDF for the observed data ($ECDF_{observed}$) and the ECDF for the simulated data ($ECDF_{simulated}$)

Step II: Calculation of the area metric (A) for each well

Step III: Calculation of the area metric (A*) for each model

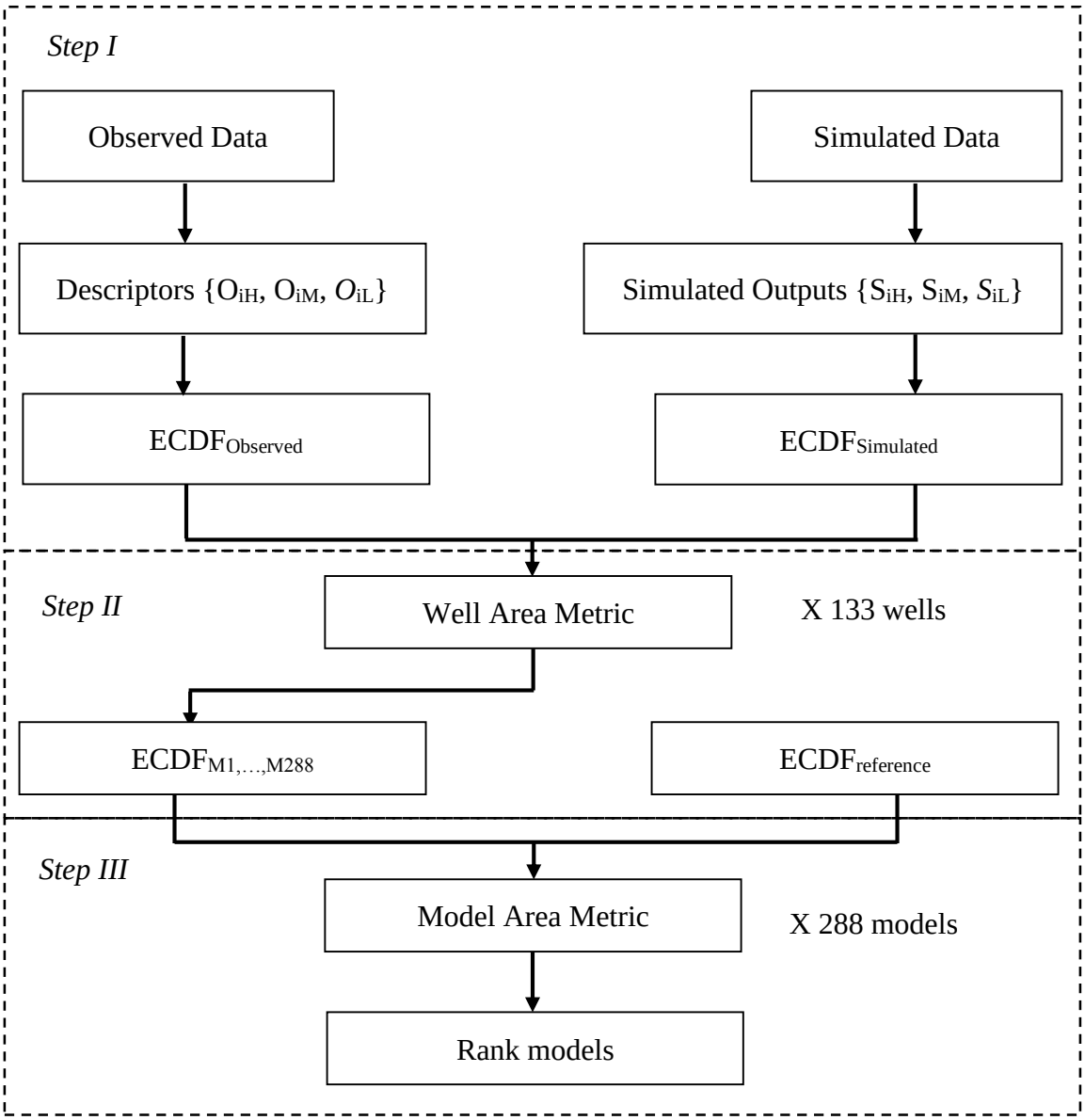


Figure 3.37: Flow chart showing the steps involved in the calculation of the area metric for wells and for the models

3.3.3.1. Step I

The first step was to generate ECDFs for the observed data and the simulated model outputs. The three steps for the observed data ECDF were derived from the maximum, the median, and the minimum head observations made at a well (O_{iH}, O_{iM}, O_{iL}). The three steps for the simulated data ECDF were derived from the high, the median, and the low hydrologic conditions (S_{iH}, S_{iM}, S_{iL}).

The groundwater head measurements made from 1976 to 2010 at 133 wells distributed across the study area were used as the observed data. These wells are screened at different depths in the underlying aquifers. The observations made at the 133 wells were used to generate 133 ECDFs, one per well. Each ECDF consisted of 3 head observations – the largest head observation, the median head observation, and the smallest head observation, collectively the "descriptors". The largest and the smallest head observations bracketed the range over which the head observations spread, while the median head observation represented the central value of this range. For instance, Figure 3.38-a shows the triangular PDF generated using the descriptors for well S3529: 22.32 feet (the smallest), 25.54 feet (the median), and 30.20 feet (the largest). Figure 3.38-b shows the corresponding ECDF.

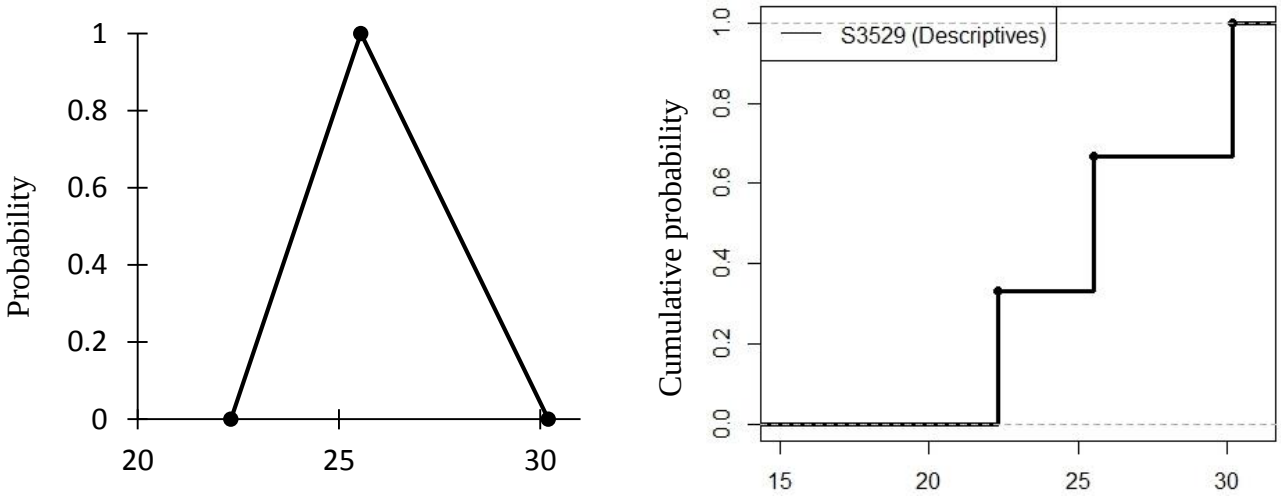


Figure 3.38: Descriptors for S3529 shown as (a) PDF and (b) ECDF

The use of descriptors for generating the $ECDF_{observed}$ was prompted because of unequal numbers of observation made at different wells. Certain wells were more consistently observed and had a longer time-frame of observations than certain other wells. For example, over a period 1975-2010, 446 head observations were made at well S3529 and 189 observations were made at well S47747, while only 30 observations were made at well S72138. Well 72165 has only 10 observations, from 1982-2010. Well S76380 has only 10 observations, from 1990-2010. Well S98434 has 68 observations, but only from 1994-2010. Therefore, the use of three descriptors ensured that all the observed data ECDFs have the same number of steps (n=3). Different $ECDF_{observed}$ had differential widths because of the variable ranges of the groundwater head descriptors. For instance, Table 3.9 shows the descriptors for 3 wells and Figure 3.39 shows the corresponding $ECDF_{observed}$. Some ECDFs appeared leaner compared to other $ECDF_{observed}$.

well	Maximum	Median	Minimum	Range
S3529	30.20	25.54	22.32	7.88
S72149	18.67	16.49	15.05	3.06
S72165	3.340	3.065	2.790	0.55

Table 3.9: Descriptors for the three head observation wells (in feet)

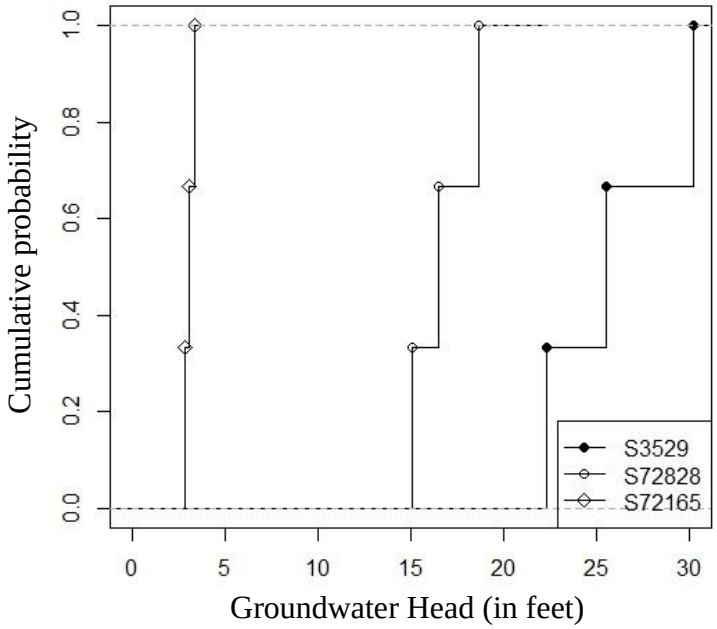


Figure 3.39: $ECDF_{observed}$ for wells in Table 3.8

The fluctuations in the head observations over the observation period represents the aleatory uncertainty associated with the observations at any given well. In addition, the epistemic uncertainty may also be associated with each of the measured heads as a result of measurement errors. Ideally, the area metric should include the measurement errors associated with the observation data (Oberkampf and Barone 2006). However, the available observational data used in this exercise included only the deterministic head observation values; no estimates of measurement error were associated with these values. Therefore, the epistemic uncertainty associated with the head measurement was not incorporated into the observed data ECDFs.

Two hundred and eighty eight model variants were simulated in three separate iterations to represent high water levels, median water levels, and low water levels. Upon simulation, the groundwater simulation models generate deterministic output quantities (simulated outputs) for each well for each states of the groundwater system (high, median, and low groundwater levels). The three simulated quantities were then arranged into an $ECDF_{simulated}$ for each model. All ECDFs are discrete, step functions with three equal-length steps. For example, Figure 3.40 shows the $ECDF_{simulated}$ of models along with the $ECDF_{observed}$ for well S3529.

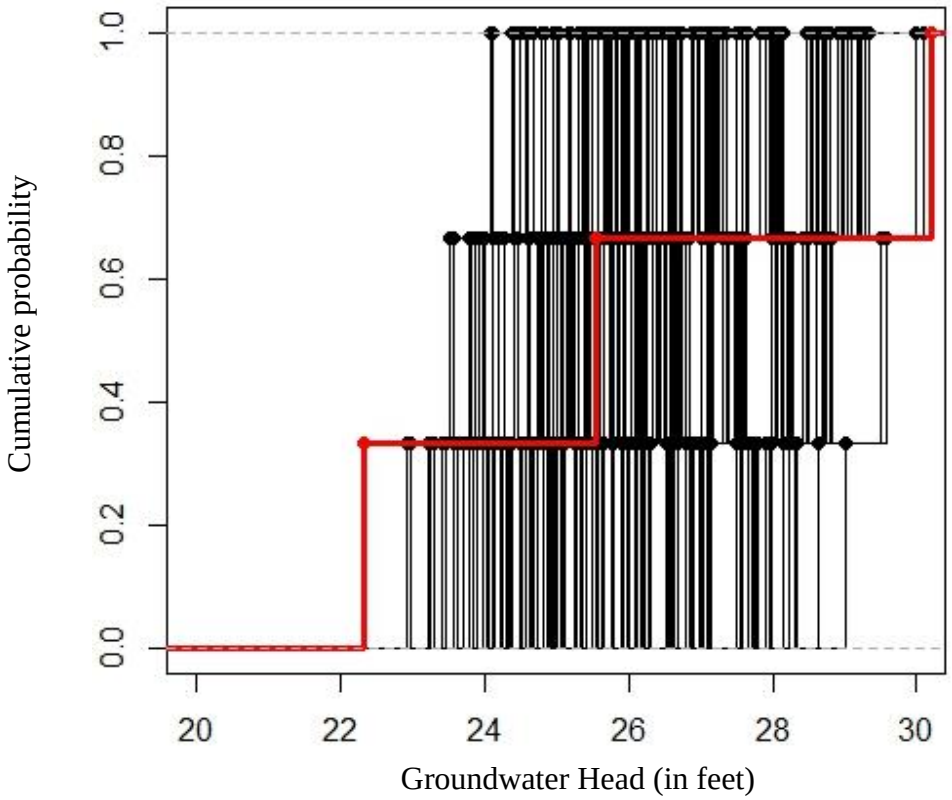


Figure 3.40: $ECDF_{observed}$ (in red) and $ECDF_{simulated}$ (in black) for the well S3529.

3.3.3.2. Step II

In the second step, the area enclosed between the $ECDF_{observed}$ and the each models' $ECDF_{simulated}$ was quantified. This process was repeated for all 133 wells, and for each of the 288 models. The following R code is used to calculate the area metric:

```
mf<-c(x,y)
ord<-order(mf);
v<-
c(rep(1/length(x),length(y)),rep(1/length(y),length(y)));
mf<-mf[ord];
v<-v[ord];
abs(sum(diff(mf)*(cumsum(v)[1:(length(v)-1)])))
print (abs(sum(diff(mf)*(cumsum(v)[1:(length(v)-1)]))))
```

(Source: https://stat.ethz.ch/pipermail/r-help/2004-February/046207.html):

Where, "x" represents the descriptors from the head values, while "y" represents the simulated head values.

3.3.3.3. Step III

The model ECDFs generated in Step II were used for calculating the model area metric (A*) for each of the 288 models. The area metric values generated for each of the 133 wells were collated into a model ECDF ($ECDF_{model}$) for each of the 288 models. Each $ECDF_{model}$ had 133 steps of each step of equal length. A* was calculated by quantifying the area enclosed between each of the $ECDF_{model}$ and the ECDF of a hypothetical reference model ($ECDF_{reference}$). The reference model was assumed to have a perfect overlap between the observed and the simulated data for each well, and as a result, the values of all well area metric value were zero (A=0, from all 133 wells). The ECDF of this reference model was a depicted as a spike distribution at zero on the horizontal axis (Figure 3.41). Smaller A* values indicate lesser disagreement between the $ECDF_{reference}$ and a given model's $ECDF_{model}$. In the figure below, the area between $ECDF_{reference}$ and $ECDF_{M1}$ is less than the area between $ECDF_{reference}$ and $ECDF_{M2}$. Then, the models were arranged in a descending order (from smallest to largest) A* values.

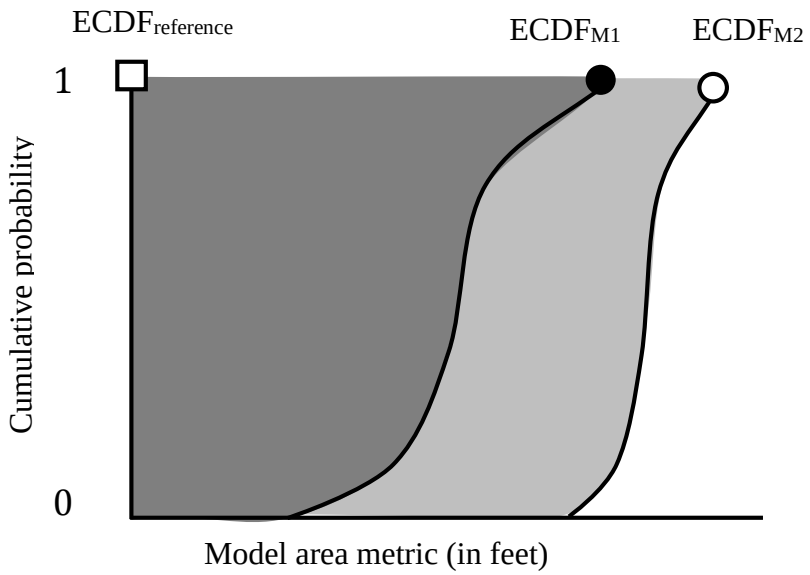


Figure 3.41: $ECDF_{M1}$ (solid circle), $ECDF_{M2}$ (hollow circle), and $ECDF_{reference}$ (open square).

Chapter 4

Results

Overview

The following discussion presents the results and the associated discussion of the multi-model validation assessment using the area-metric method for the Brookhaven landfill case study groundwater flow simulation model. Twenty-three models were excluded from the analysis due to abnormal termination of their simulations. The remaining 265 model ECDFs were subjected to the analysis of logical ordering where 200 models were retained for further analyses; other models were rejected due to the violation of logical ordering. Descriptive statistics associated with the model area metric values (A*) were discussed. The top 10 models with the lowest A* values, along with the bottom 10 models with the highest A* values, were selected and their configurations were analyzed. This was followed by the analysis of model performances at each of the 133 wells. Descriptive statistics and spatial distribution of the well area metric values (A) were analyzed and discussed; the A values of the top 10 models, and that of the whole model set, were compared. The A* values were classified into the states of the variable features that were then subjected to a qualitative boxplot analysis and a one-way ANOVA. The association between the A* values and the performance of the models measured using the traditional RMSE metric was analyzed. Finally, sensitivity of the results to inclusion/exclusion of model features, to the change in the descriptor values, and to the change in the resolution of the well ECDFs were analyzed and discussed.

The results of the area metric-based multi-model validation assessment for the Brookhaven landfill case study groundwater flow simulation model were as follows.

4.1. All Models

A total of 288 model variants were simulated and evaluated. Of these, 265 models produced outputs, while 23 models' simulations were abnormally terminated. The $ECDF_{simulated}$ were compared with the corresponding $ECDF_{observed}$ for each of the 133 wells; this produced 133 well area metric values (A values) for each of the 265 models. The 133 A values generated for a given model were used to derive the $ECDF_{model}$ for that model. A total of 265 $ECDF_{model}$ were generated; these model ECDFs were graphed (Figure 4.1). The figure also shows the $ECDF_{reference}$ where A = 0 feet for all wells.

The figure shows intertwining of the model ECDFs. The model ECDFs occupied different positions and had variable spreads. The probability that the A values for any well is over 0 feet exceeded zero ($p > 0$) for all models, while the probability that the A values for any well was under 8 feet was one ($p = 1.0$) for all models. The spread of the ECDFs was highest in wells with larger A values (located at or close to probability 1.0), while the spread of the ECDFs was comparatively compact in wells with smaller A values (located at or close to probability 0). This was interpreted as model result inconsistency: models that performed better with regard to some wells did not perform better than all models with respect to other wells.

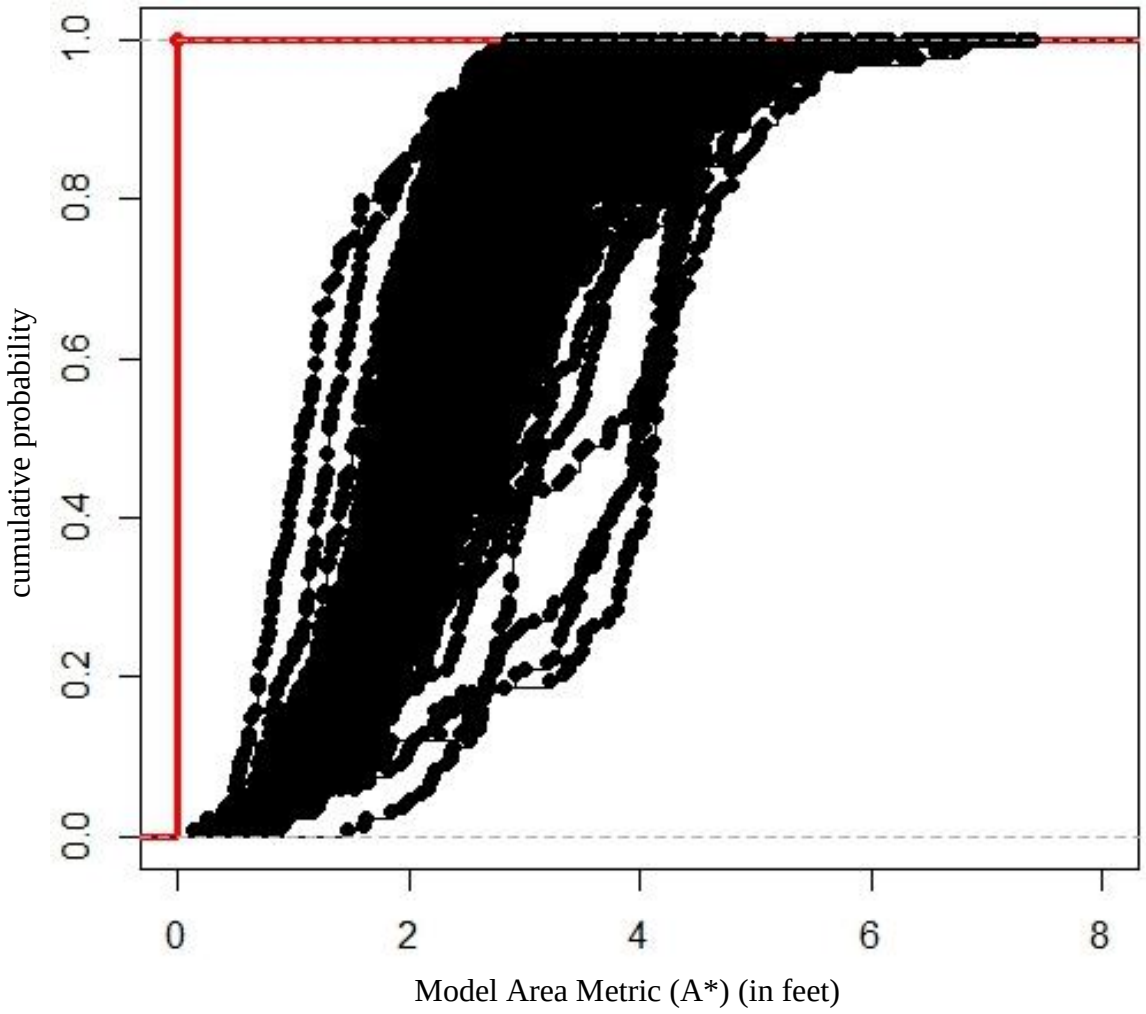


Figure 4.1: $ECDF_{model}$ (in black) and the $ECDF_{reference}$ (in red)

The abnormal termination of 23 models occurred typically because the model simulation failed to meet the convergence criterion within the designated number of inner and outer model iterations. Abnormal termination occurred in models simulating low water table conditions with the bottom of the first layer (L1) higher than the elevation of the northern CHD boundary in some grid cells. This combination resulted in dry cells in the top layer and the program indicated error in the simulation. Changing the model settings, increasing the number of inner or outer iterations, changing the solver for MODFLOW-2000 from PCG to WHS, and modifying the height of the bottom of layer 1 resulted in successful completion of simulation for 7 out of the 23 terminated models. However, changes in the individual model settings were not accommodated since these model features were kept fixed for all models. Consequently, the abnormally terminated models were excluded from further analyses.

4.2. Analyses of Logical Ordering

For each well, the $ECDF_{observation}$ was always monotonic increasing and composed of three data points: the minimum, median, and maximum groundwater head values (see Figure 4.2). Similarly, the $ECDF_{simulated}$ was also composed of three data points, resulting from simulating three states of the groundwater system: low groundwater conditions, the median groundwater conditions, and the high groundwater conditions. These data points were always arranged to create a monotonic increasing ECDF for each of the 288 models. It is expected that the model simulations follow a logical ordering, that is, that the model simulating high groundwater level state should generate head values larger than the head values simulated by the model simulating median groundwater level state (those, in turn, should be larger than the head values simulated by the model simulating low groundwater level state). For this logical ordering to hold true, the difference between the high and the median, and the median and the low simulated head values should always be positive. A model that does not reproduce this pattern seems to violate basic expectations.

Head Observation (in feet)	Minimum	Median	Maximum
	22.32	25.55	30.20

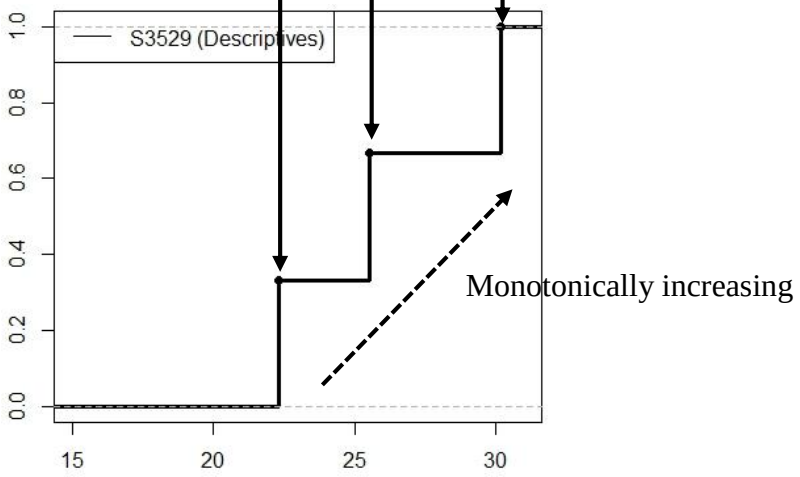


Figure 4.2: $ECDF_{observed}$ for the well S3529

Therefore, models were rejected as violating logical ordering when the difference "High - Median" and "Median - Low" was greater than -0.02 feet for more than 10% of the wells (14 or more of the 133 wells); this rejection criterion and the percentage of wells were chosen so that more than just a few violations of the precept were needed to identify a "faulty" model. For example, Table 4.1 shows results from two models, illustrated by data from well S3529. Although, the A* value in model M19 was very similar to the A* value for model M1, the difference between the median and the low simulated values indicated violation of logical ordering in M19. This condition was measured at all 133 wells for this model and therefore this model was rejected. Similar violations of logical ordering occurred in only one well for model M1.

Model	A* (feet)	Simulated Heads (feet)			High – Median	Median – Low	Rejected (Y / N)?
		High	Median	Low			
M1	1.12	25.94	25.38	24.83	0.56	0.55	N
M19	1.08	25.36	24.81	27.58	0.56	-2.77	Y

Table 4.1: Investigating violation of logical ordering in well S3529 for models M1 and M19

In the area metric approach, the temporal ordering of data is disregarded in the ECDFs. That is, the data values are re-arranged in an increasing order of their magnitudes (from smallest to largest), independent of their chronological ordering, to develop the positively monotonic ECDFs. As a result, the correspondence between particular observed values and simulated values is lost and this loss may not be recognized if the model simulates right values for the wrong reasons. For example, a model intended to simulate low groundwater level may generate head values that are larger than the outputs produced by the same model simulating high groundwater levels. This logical fallacy is not obvious because of the arrangement of data on the basis of magnitude. It is suggested that the non-randomness among the constituents of distributions, and the structured dependence between observed data values and simulated data values should be separately explored (Ferson et al. 2008). Simple subtraction of the simulated values was applied for this exploration.

Sixty-five of the 265 models that remained were rejected due to the violation of logical orderings and 200 models were retained for further analysis. Figure 4.3 shows the ECDFs of the 200 model variants that were retained, along with the $ECDF_{reference}$.

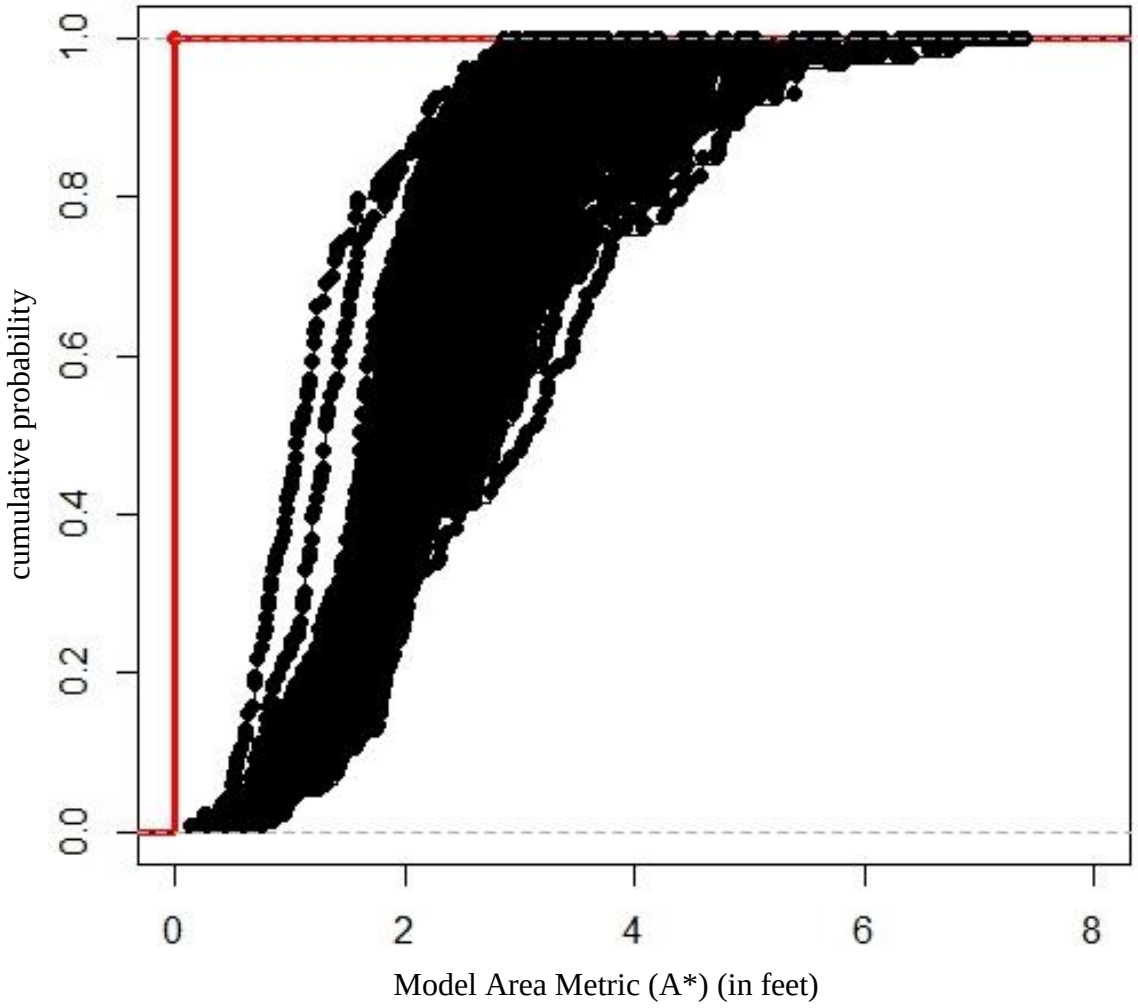


Figure 4.3: $ECDF_{model}$ of 200 model variants (black) with the $ECDF_{reference}$ (red)

4.3. Model Area Metric (A*)

The values of the model area metric (A*) were calculated as the area enclosed between the $ECDF_{reference}$ and each of the 200 $ECDF_{model}$. The A* values from the models were arranged in a CDF (Figure 4.4). The probability of A*< 2 feet was about 0.2 ($p \simeq 0.2$), while the probability of A*< 3 feet was one ($p = 1.0$).

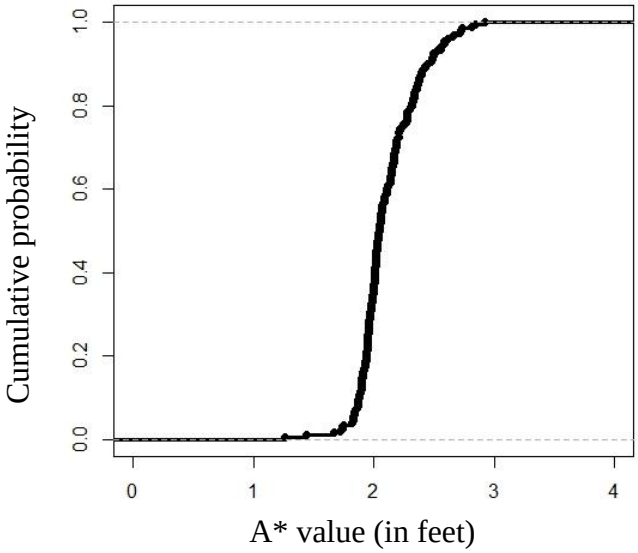


Figure 4.4: A* values of the 200 models as a CDF

The descriptive statistics associated with these A* values are shown in Table 4.2. Model #178 achieved the smallest A* value; therefore, model # 178 was considered to be the most representative of the groundwater flow regime in the vicinity of the Brookhaven landfill from the model space of 288 models, on the basis of the assessment of the replicative validity of these models.

Description	A* value (in feet)
Minimum	1.25 feet (model #178)
Maximum	2.92 feet (model #169)
Median	2.04 feet
Mean	2.11 feet
Standard deviation	0.24 feet

Table 4.2: Descriptive statistics associated with the A* values of 200 models

Notwithstanding the disjoint configurations of the constituent models, one can create an ensemble from models from the above models. However, a model ensemble with a larger spread of goodness-of-fit would include high skilled models as well as comparatively lesser skilled models. In this way, the proposed approach can be used in improving the composite goodness-of-fit of an ensemble by ranking the constituent models on the basis of their area metric scores. Models above certain rank threshold such that only those models with better skills would be included, narrowing the spread of the ensemble.

Figure 4.5 shows box and whiskers plots depicting the minimum, maximum, medians, interquartile ranges, and outliers of the 133 A values calculated for each of the 200 models. The boxplots were arranged in an ascending order of their A* values (superimposed on the boxplots in red). The A* values gradually increase from left to right, from about 1.25 feet to about ~ 3 feet. The A* values of the first five models was distinctly low relative to a steady increase in A* values in the remaining models. The medians showed a near identical pattern with the A* values. The interquartile ranges increased from 0.52 feet (model #189) to 2.57 feet (model #169). The spread of the interquartile as well as total ranges of the A values is small for models with smaller A* values, while the spread gradually increases left to right with the increase in the A* values. The magnitudes of the A* value appeared to increase with an increase in the total and the interquartile ranges; the increase in the ranges was gradual for the first 100 models in the figure, and the increase is more noticeable thereafter. A number of singular outliers extended beyond the interquartile range and were observed for most of the models. The correspondence between the A* values, the total ranges and the interquartile ranges underlined the notion that the area metric depicts the differences not only in the central tendencies, such as the means, but also the differences at the tails of the distributions that represent the extreme values. Larger A* values were generated for those model ECDFs that had a wider spread compared to those models with comparatively smaller spread.

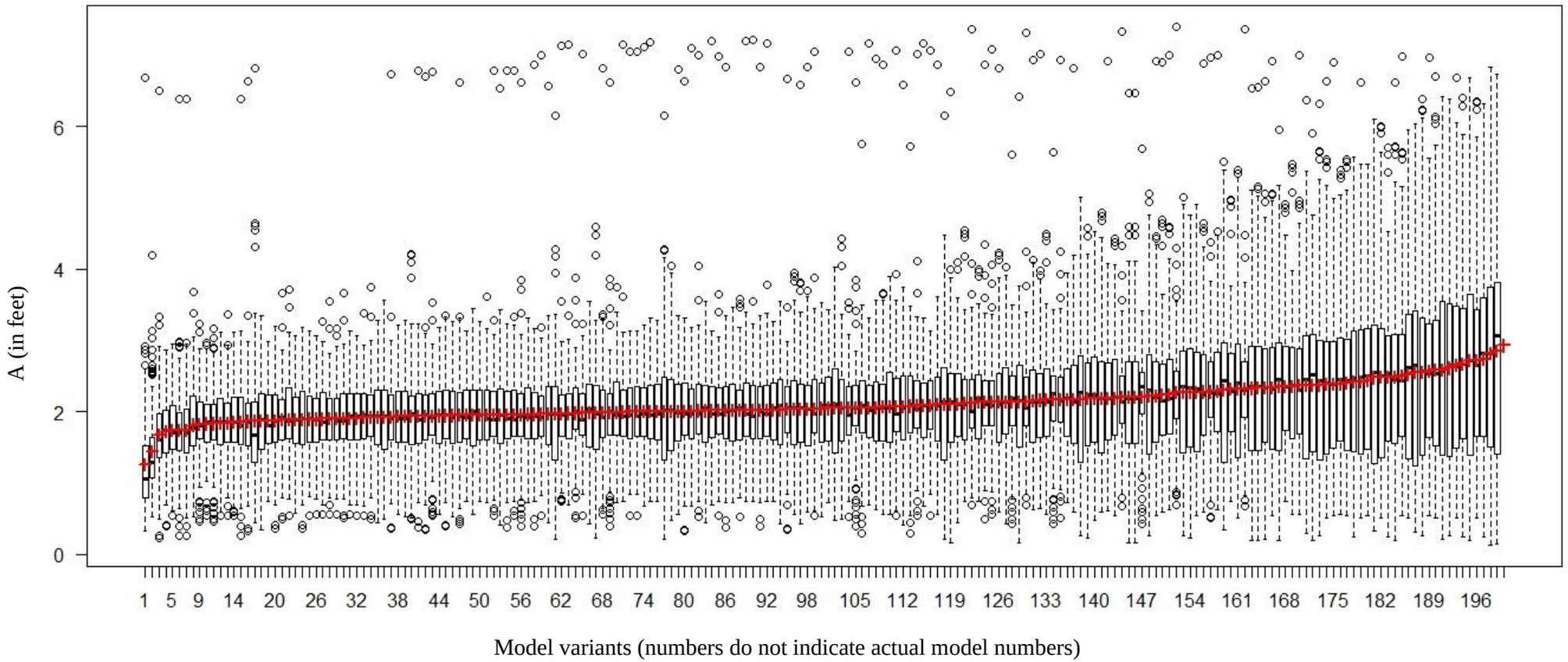


Figure 4.5: Boxplots showing ascending arrangement and the range of the A values for each of the 200 models; the corresponding A* values are superimposed (in red)

The 200 models were arranged from the smallest to the largest on the basis of their A* values. Figure 4.6 shows color-coded the $ECDF_{model}$ of the top 10 models that appeared among the closest to the $ECDF_{reference}$. This figure qualitatively shows that the top 10 models were among the best representations of the groundwater regime in the Brookhaven landfill vicinity, from the model space of 288 models, on the basis of their replicative validity that was assessed using the area metric approach. As mentioned above, the model performance varied within each $ECDF_{model}$, including the $ECDF_{model}$ for the top 10 models, given the dispersion seen within each ECDF; dispersion appeared to be more pronounced near p=1.0 where comparatively larger A values could be seen.

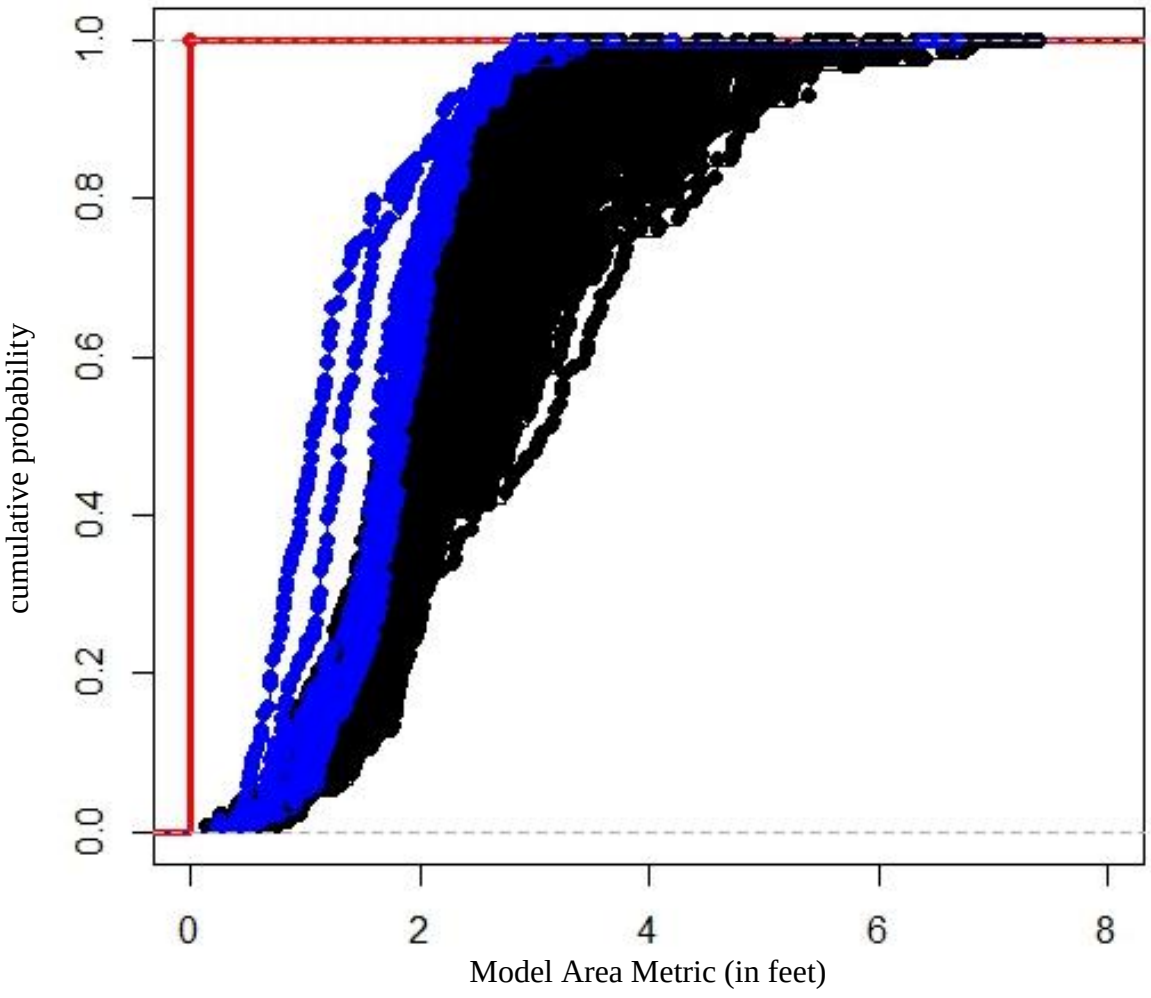


Figure 4.6: Top 10 models' $ECDF_{model}$ (in blue), other $ECDF_{model}$ (in black), and $ECDF_{reference}$ (in red)

Figure 4.7 shows the mean of the A values calculated from the A values of the top 10 models, the bottom 10 models with the largest A* values, and the mean of the A values calculated from the A values of all 200 models (including the top 10 models) (overall mean A values). The figure indicated that the means of the A values from the top 10 models are generally lower than the means of the A values from all models; the difference these two values was less than zero (net negative) for 119 out of 133 wells. The means of the A values from the bottom 10 models were generally higher than the means of the A values from all models; the difference these two values net negative for 40 out of 133 wells. The differences were more pronounced for wells that had larger overall mean A values, while the differences were comparatively smaller for wells with smaller overall mean A values. This indicated that the top 10 models, as assessed by the area metric, performed better than the model set as a whole.

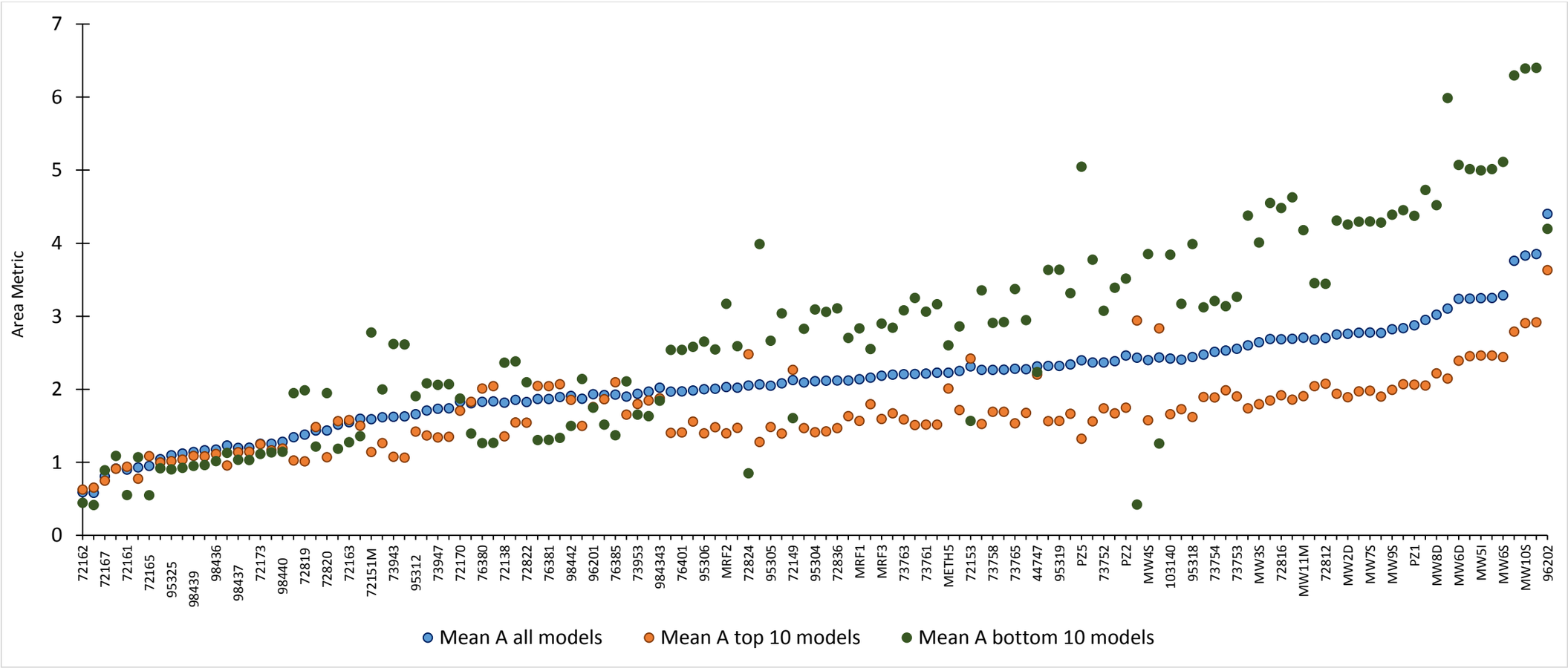


Figure 4.7: Means A values for the whole model set, for the top 10 models, and for the bottom 10 models

Figure 4.8 below shows the boxplot of the mean A values indicate that the median of the mean A values for the top 10 models is lower than that for all models that, in turn, is lower than that for bottom 10 models.

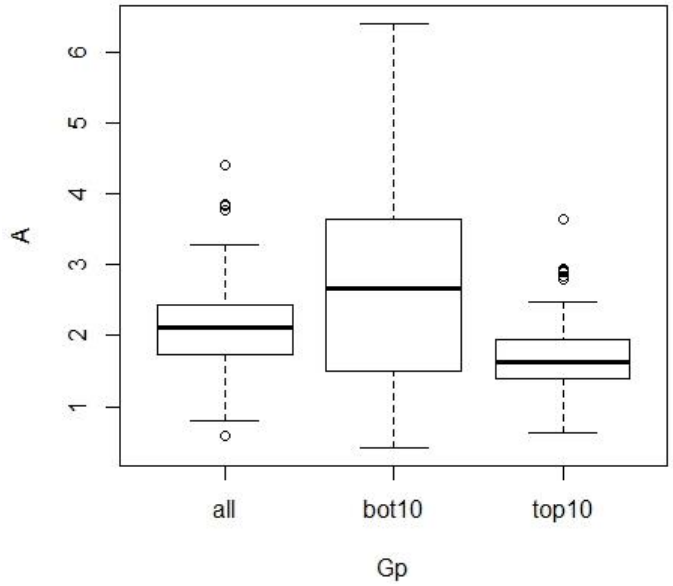


Figure 4.8: Boxplot of the mean A values

The differences in the mean A values were tested for statistical significance using ANOVA. The test shows that there is significance difference in means between the mean A values among different model sets ($F = 41.92$, $p = <2e\text{-}16$). Pairwise t-tests (with Bonferroni p-adjustment) suggests that the means were significantly different between the all model set and the bottom 10 models (p-value = 4.5e-07), between all model set and the top 10 models (p-value = 0.00058), and between the top 10 models and the bottom model set (p-value = < 2e-16).

4.4. Well Area Metric (A)

Well area metric values (A) were generated for the 133 wells for a given model. These 133 A values for a given model were subsequently used to generate the model ECDFs and subsequently to calculate A* value for that model. Patterns and geospatial distributions of these A values were explored, as in Figure 4.9 which shows the 200 $ECDF_{simulated}$ for well S3529 along with the $ECDF_{observed}$. The A value was calculated for the difference between the $ECDF_{observed}$ and each of the 200 $ECDF_{simulated}$.

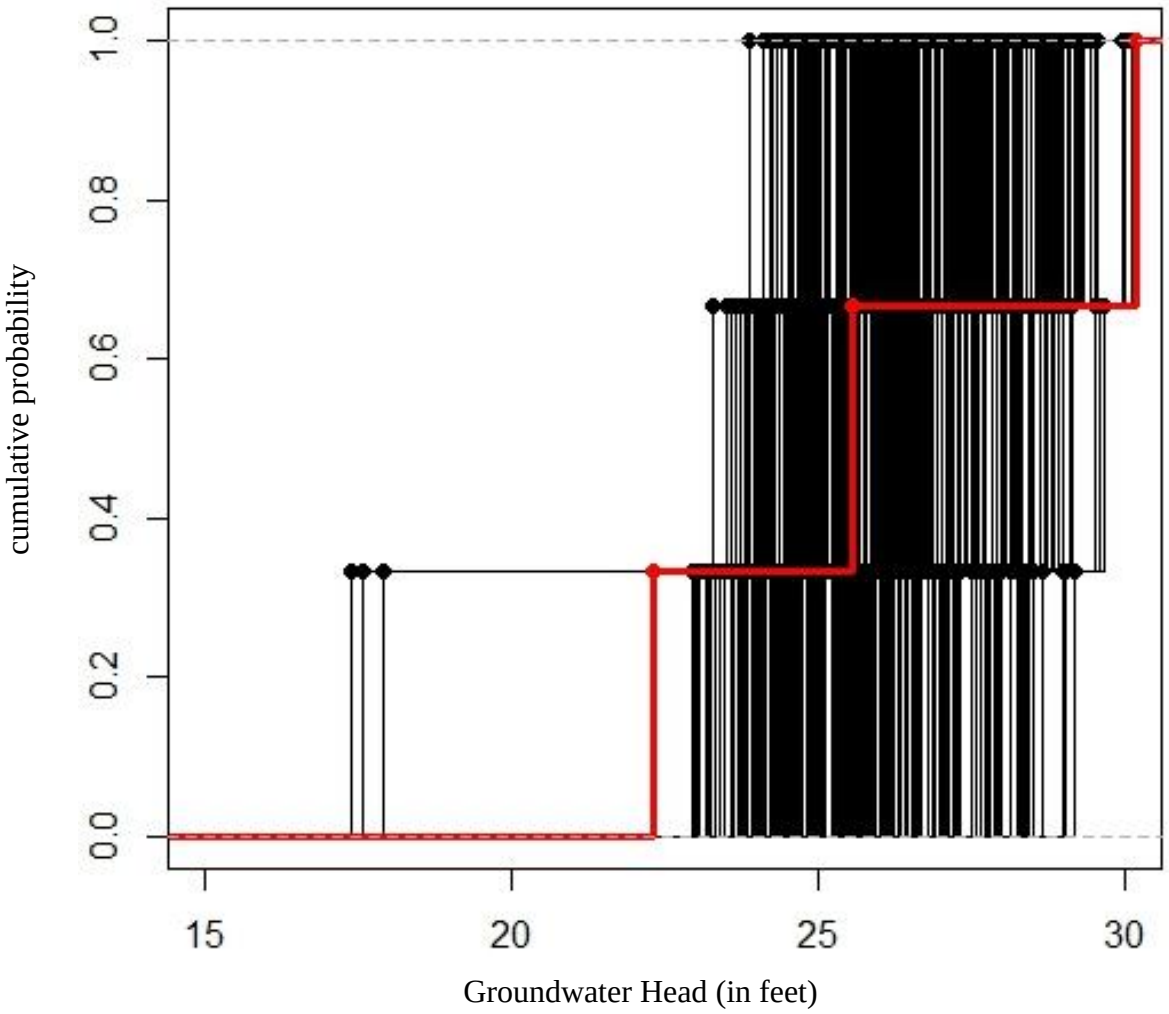


Figure 4.9: Model $ECDF_{simulated}$ (black) and $ECDF_{observed}$ (red) for the well S3529

Figure 4.10 shows the descriptors – mean, one standard deviation, minimum, and maximum – of each of the 133 A values arranged in an increasing order. Generally, the range (difference between the maximum and the minimum A value) increased with the increase in the mean A values. The standard deviation associated with the mean values did not follow this pattern; smaller spread was observed in wells with larger mean A values and larger standard deviation was observed in wells with comparatively smaller mean A values. The range of A values indicated that the simulated outputs across the 200 model variants varied considerably in response to the changes in model configurations.

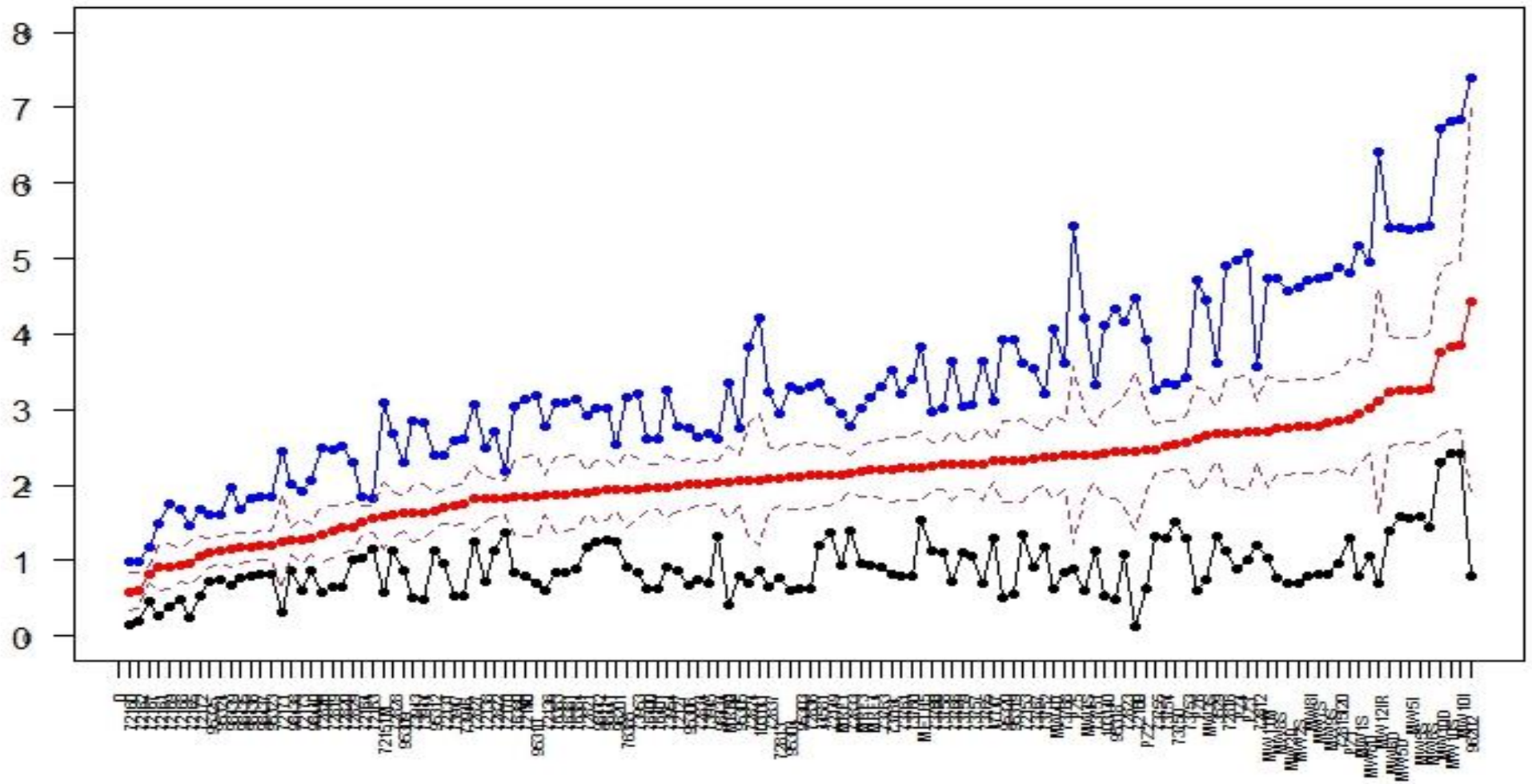


Figure 4.10: The mean (red), one standard deviation (dotted lines), minimum (black), and maximum (blue) A values for the 133 wells

Table 4.3 shows the lower and upper limits on the descriptors of the A values along with the wells where these limits were observed, as well as the count of observations made at these wells.

Descriptors	Lower limit: A value (well)(count)	Upper limit: A value (well)(count)
Maximum	1.64 (S72165) (2)	7.40 (S96202) (66)
Minimum	0.13 (S72159) (2)	2.43 (MW10-I) (108)
Mean	0.60 (S72162) (4)	4.26 (S96202) (66)
Std Dev	0.16 (S72167) (2)	2.57 (S96202) (66)
Range	0.78 (S72163) (5)	6.60 (S96202) (66)
1^{st} percentile	0.40 (S72162) (2)	3.02 (MW10-I) (108)
3^{rd} percentile	0.76 (S72162) (2)	6.83 (S96202) (66)

Table 4.3: Lower and upper limits on the descriptors of A values (in feet)

The wells that constituted the lower limit A values (S72159, S72162, S72163, S72165, S72167) had distinctly low sample counts than those wells that constituted the upper limit A values (S96202, MW-10I). However, the difference in the counts and its influence on the A values was neutralized by using three data points to formulate the $ECDF_{observation}$ for each well. Descriptors for well S96202 were derived from observations made during periods 1995-2001 and 2009-2010, while the descriptors for the well MW10-I were derived from period 1993-2002. Descriptors for wells S72159, S72162, S72163, S72165, and S72167 were derived from observations made during 2010. Therefore, S96202 and MW-10I supposedly represented a bigger period compared to the lower limit group of wells that supposedly represented comparatively more recent periods. Thus, the discrepancies in the A values between these two groups of wells may have been a function of the period from which the descriptors were derived.

Well S96202 had the highest mean A value (4.26 feet) with highest spread (one standard deviation of about 2.57 feet), the largest maximum values (7.043 feet), and the largest range (6.602 feet). This well was screened at about -125 feet msl in the shallow Magothy aquifer in each model variant. Given the large A values, this depiction of the screen zone could be wrong. One plausible alternative could be that the screen is located in the PSU and that the Magothy aquifer is deeper than where it is depicted in the present models.

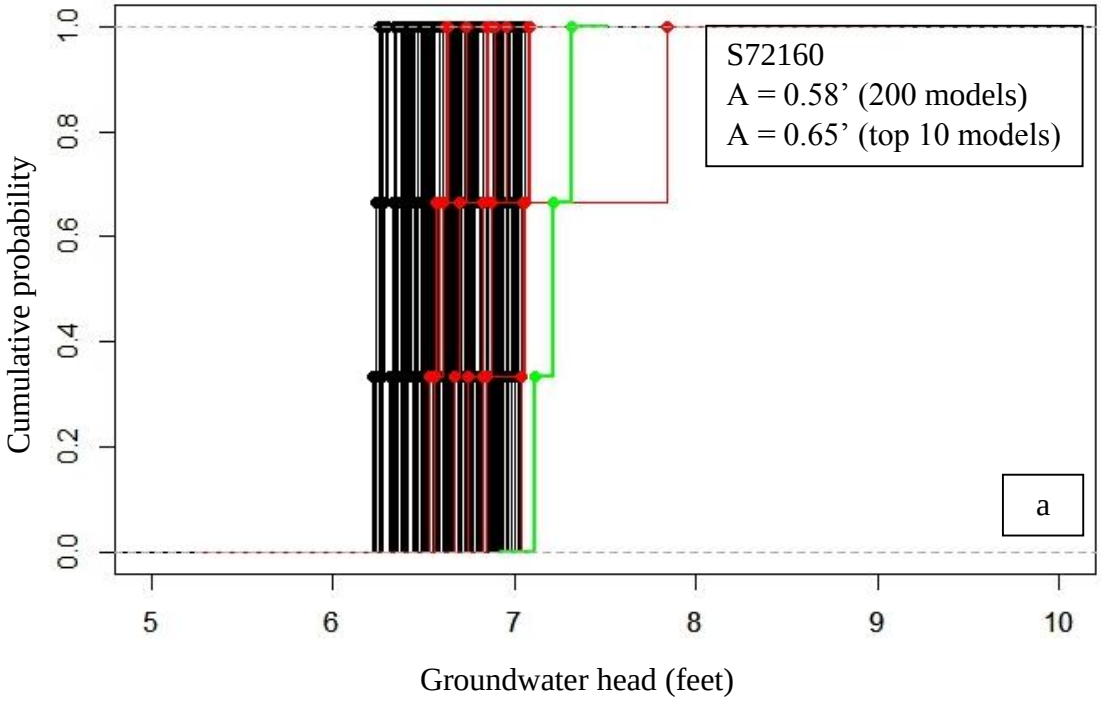
Area metric values are dependent on the quality of the data; if the same number of observations were taken during the same periods, it could eliminate the chronological bias in the area metric values. These conditions may not be encountered in every assessment; adjustments may be needed. In such cases, it is important to qualify the results of the area metric with the adjustments being made. Nonetheless, a detailed inspection of the A values can reveal potential shortcomings of the hydrogeologic configuration of the model domain and could be used to decide the direction of changes in the configuration.

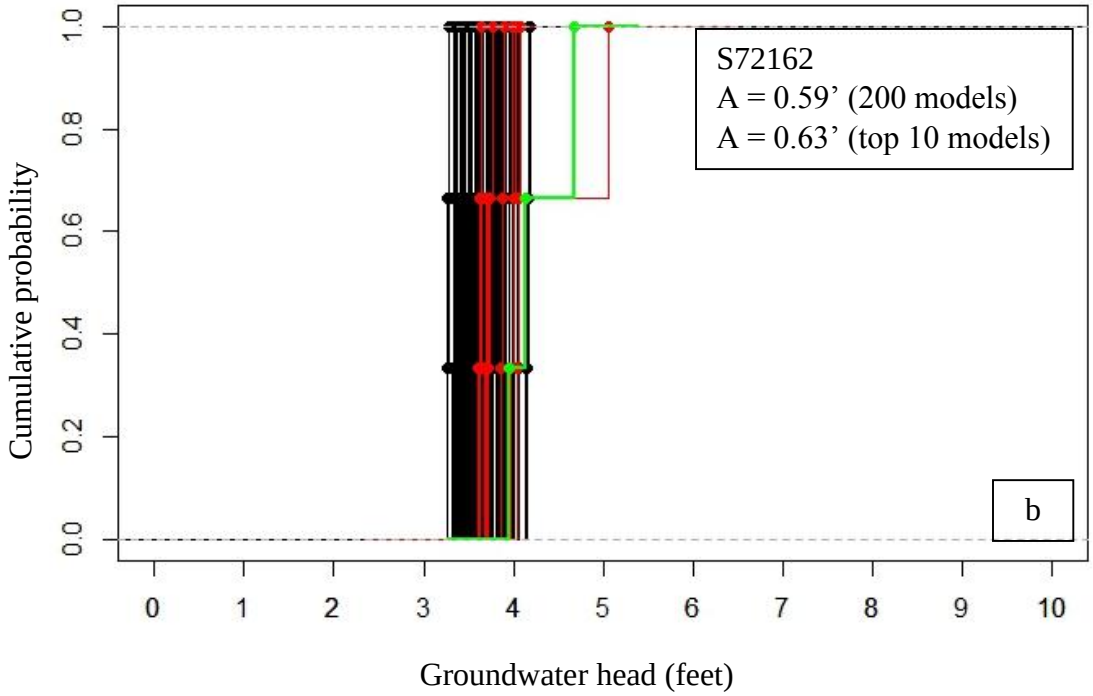
Well MW10-I had the highest of the minimum A values (2.43 feet) as well as the highest 1st percentile A value (3.02 feet). This well is part of a triplet of wells located between Cell 2 and Cell 5, at the southwestern corner of the old landfill. Well MW10-I represents the well screened at the intermediate depth (-20 feet msl) into the Upper Glacial aquifer; MW10-S is screened at shallow depth (11 feet msl), while MW10-D is screened the deepest (-78 feet msl). All three wells showed similar descriptor A values (Table 4.4). This suggested that the area metric values resonated similarities in the well locations. The location of the wells dominated over the variation in the screen depths while deciding their corresponding A values. Generally, the head values at these wells are near identical to each other; therefore, these A values are consistent with the pattern of observed head values at these wells. This cluster was abandoned when Cell 6 was constructed.

Descriptors	MW10-S	MW10-I	MW10-D
Maximum	6.81	6.84	6.73
Minimum	2.41	2.42	2.31
Mean	3.82	3.84	3.76
Std Dev	1.07	1.07	1.06
Range	4.40	4.41	4.42
1st percentile	2.99	3.02	2.94
3rd percentile	4.56	4.56	4.46

Table 4.4: Descriptor A values for MW-10S, MW-10I, and MW-10D

As mentioned in Figure 4.6, the mean of the A values of the top 10 models were smaller than the mean of the A values derived from the entire model set of 200 models in 119 out of 133 wells. Figure 4.11-a-d show the $ECDF_{observed}$, $ECDF_{simulated}$ for the top 10 models, and the $ECDF_{simulated}$ for the remainder of the 190 models for wells S72160, S72162, MW10-I and 96202. It can be seen that the $ECDF_{simulated}$ for the top 10 models are generally closer to the $ECDF_{observed}$ in wells with small A values (S72160, S72162), while the $ECDF_{simulated}$ for the top 10 models are thinly spread across the distributional range, farther away from the $ECDF_{observed}$ in wells with larger A values (S96202, MW10-I).


S72160
A = 0.58’ (200 models)
A = 0.65’ (top 10 models)
a
Cumulative probability
0.0
0.2
0.4
0.6
0.8
1.0
5
6
7
8
9
10
Groundwater head (feet)


S72162
A = 0.59’ (200 models)
A = 0.63’ (top 10 models)
b
Cumulative probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
8
9
10
Groundwater head (feet)

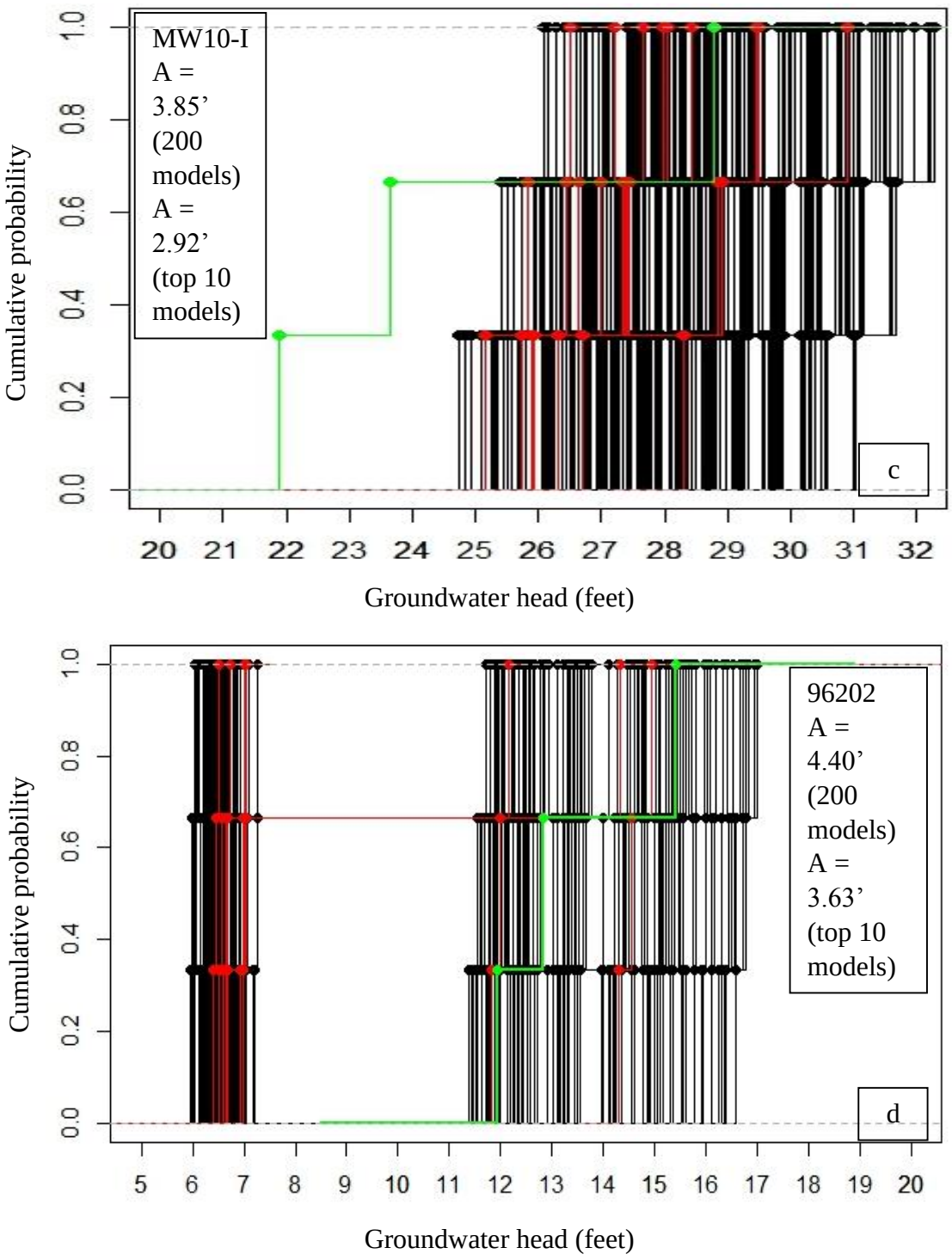


Figure 4.11: $ECDF_{observed}$ (green), $ECDF_{simulated}$ for top 10 models (red), and $ECDF_{simulated}$ for the remainder of the 190 models (red) for (a) well S72160, (b) well S72162, (c) well 96202, and (d) well MW10-I

The 133 wells were distributed across the model domain and were screened at different depths. Wells were clustered between the northern perimeter of the landfill boundary and the downgradient region between Beaverdam Creek and Little Neck Run. Wells were relatively scarce in other parts of the model domain. The geospatial distribution of the mean A values was spatially visualized in the form of color-coded solid circles ("bubbles") where the size of the bubbles was directly proportional to the magnitude of the mean A values (Figure 4.12). The distribution of the mean A values indicated that larger values were near the northern and the southern edge of the landfill and in the upper reaches of the Beaverdam Creek located downgradient of the landfill. The mean A values were smaller in areas further downgradient. This also visualized the discrepancy between the upper and the lower limits of the descriptors mentioned in Table 4.4 above; the wells constituted the lower limits on the descriptors were located in the southern portion of the model domain.

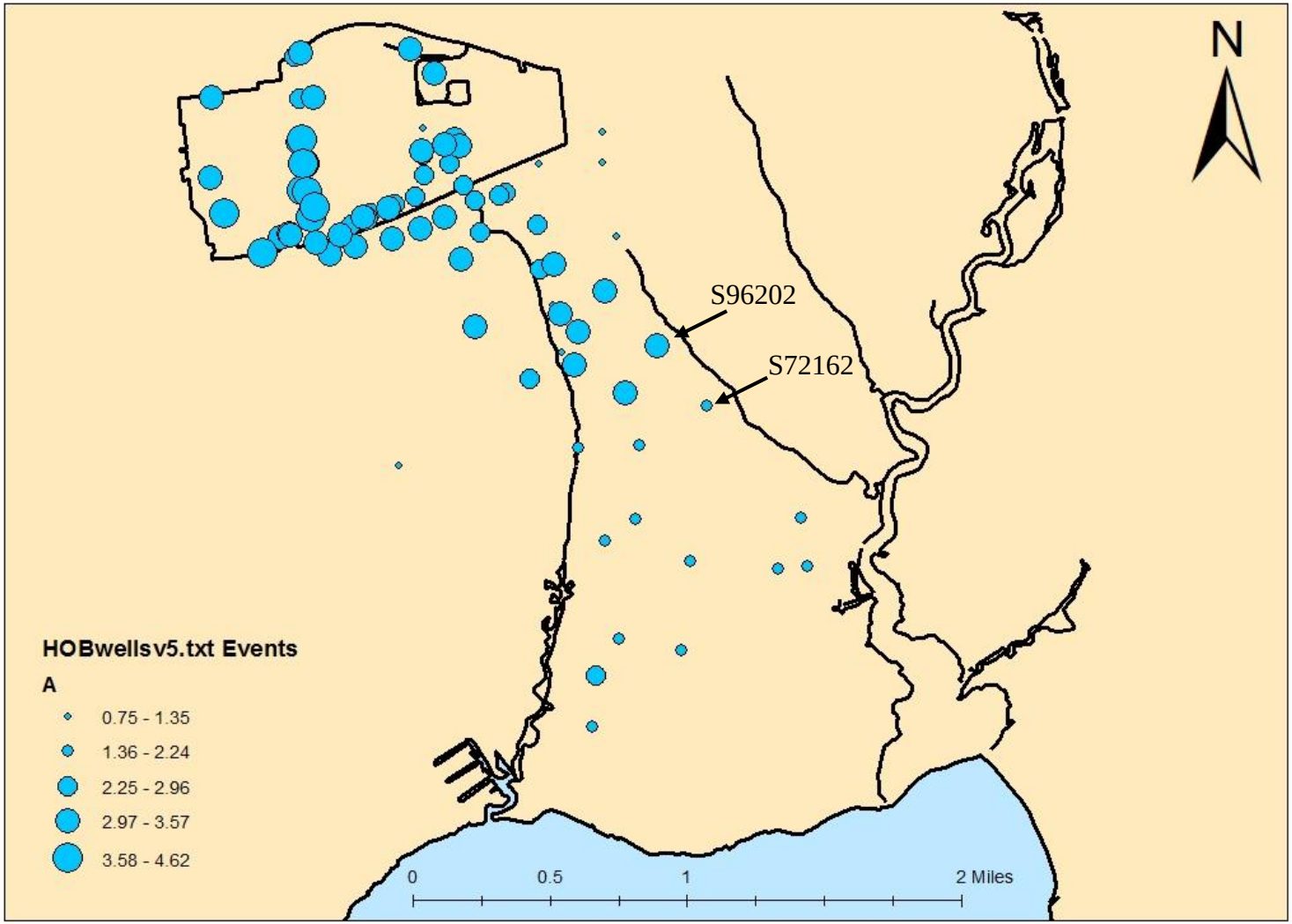
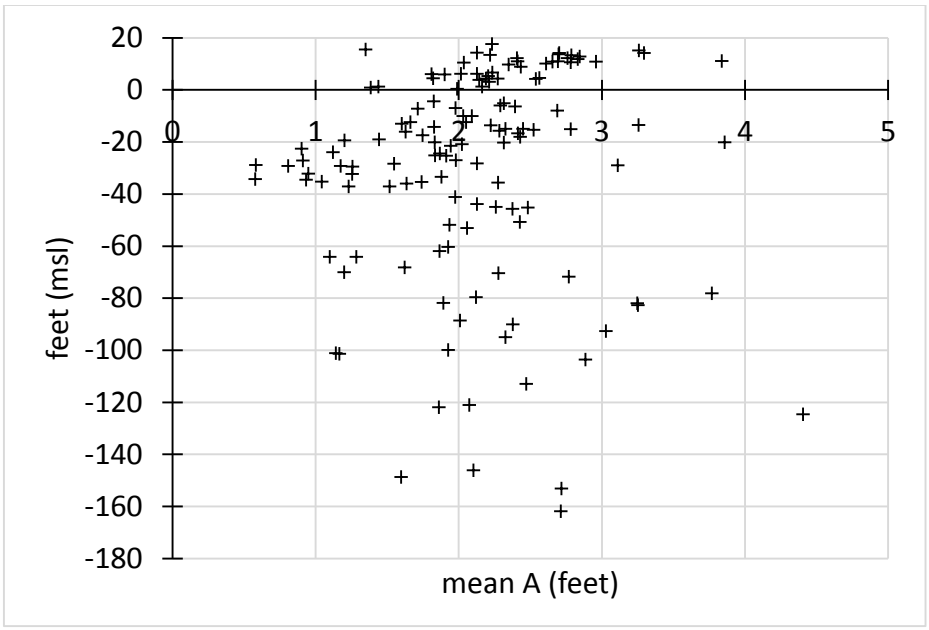


Figure 4.12: Spatial distribution of the mean A values

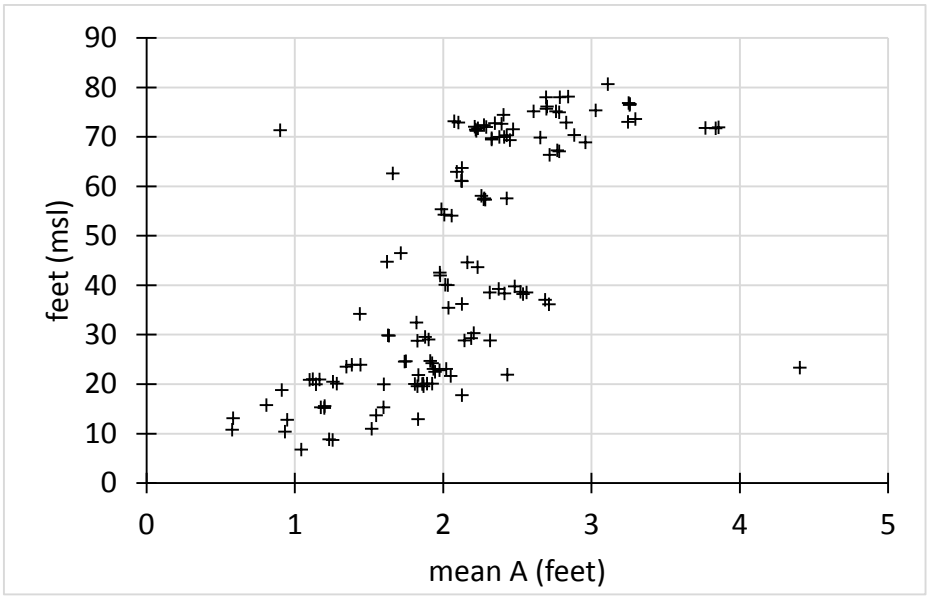
The vertical distribution of the A values was also analyzed. Figure 4.13-a shows the vertical distribution of the mean A values with respect to the screen depth. The well density was higher between 20 feet msl to -40 feet msl, while the distribution of deeper wells is less robust. The mean A values with respect to screen depth were spread between 0 and 3 feet. The A values for the wells with screen depths below -50 feet ranged mostly from 1 feet to 4 feet. An outlier data point, S96202, is visible at around -120 feet msl with the mean A value of 4.40 feet.

Figure 4.13-b shows the vertical distribution of the mean A values with respect to the measuring point elevation, the point of known elevation at or above the ground surface from which the groundwater heads are measured. The vertical distribution of the means A values indicated that the values increased with the increase in the measuring point elevation. An outlier data point, S96202, is visible at around 30 feet msl with the mean A value of 4.40 feet.

A total of 40 of the 133 wells were screened between 20 feet and 0 feet (the shallow zone), 62 wells were screened between 0 feet and -50 feet (the intermediate zone), and 31 wells were screened at depths below -50 feet (the deep zone). Among the wells screened in the deep zone, 4 wells were screened in the shallow Magothy aquifer (-140 feet to -160 feet) while 2 wells (S96202 and S95310) were supposedly screened in the potentially semi-confining unit (PSU) (-120 feet to -125 feet). The well density was relatively higher in the shallow zone, in the Upper Glacial aquifer. Therefore, these sections of the model domain were more favorably represented in the models' validation assessment compared to other areas. This spatial bias may have been transferred on to the calculation of the A* values. The topography of the study site falls from northwest to southeast. Therefore, the bias seen in Figure 4.12 is similarly seen in Figure 4.13-b.



(a)



(b)

Figure 4.13: Vertical distribution of the mean A values (a) with respect to the depth of the screen zone of the wells, and (b) with respect to the measuring point of the wells (in feet)

4.5. Analyses for Variable Features

The model configurations, that is, the combinations of the states of the uncertain variable features for the top 10 and the bottom 10 modes are shown in the Table 4.5 below.

Model		A*	Configuration (variable states)						
Top 10 models	178	1.258	V12	V21	V32	V43	V51	V61	V72
	265	1.441	V12	V21	V31	V41	V51	V61	V72
	200	1.667	V12	V22	V31	V41	V52	V63	V72
	244	1.717	V12	V22	V32	V42	V52	V62	V71
	177	1.738	V12	V21	V32	V43	V51	V61	V71
	204	1.740	V12	V22	V31	V42	V51	V62	V72
	216	1.745	V12	V22	V31	V43	V51	V62	V72
	242	1.814	V12	V22	V32	V42	V52	V61	V71
	213	1.827	V12	V22	V31	V43	V51	V61	V71
	141	1.830	V12	V21	V31	V43	V51	V61	V71
Bottom 10 models	42	2.588	V11	V21	V32	V41	V51	V63	V72
	181	2.611	V12	V21	V32	V43	V51	V63	V71
	170	2.652	V12	V21	V32	V42	V51	V63	V72
	53	2.662	V11	V21	V32	V42	V51	V63	V72
	163	2.718	V12	V21	V32	V41	V52	V63	V71
	157	2.724	V12	V21	V32	V41	V51	V63	V71
	47	2.736	V11	V21	V32	V41	V52	V63	V71
	175	2.813	V12	V21	V32	V42	V52	V63	V71
	41	2.846	V11	V21	V32	V41	V51	V63	V71
	169	2.926	V12	V21	V32	V42	V51	V63	V71

Table 4.5: Configurations of the top 10 models (with the smallest A* values) and the bottom 10 models (with the largest A* values)

The top 10 and the bottom 10 models were classified on the basis of the particular feature-states their configurations contain in the Table 4.6 below. For example, all top 10 models contained state V12 (variable thickness of the bottom of L1); no top models contained state V11

(uniform thickness of the bottom of L1). On the other hand, 4 of the bottom 10 models contained state V11, while the remainder 6 models contained state V22.

Code	Variable Feature	State 1	State 2	State 3
V1	Bottom of layer 1	Uniformly thick **(0)** (4)	Variably thick **(10)** (6)	--
V2	Bottom of layer 2	Uniformly thick **(4)** (10)	Variably thick **(6)** (0)	--
V3	Extent of the PSU	2-zone **(6)** (0)	3-zone **(4)** (10)	--
V4	Recharge	Natural **(2)** (5)	Via Recharge Basins **(3)** (4)	No recharge **(5)** (1)
V5	Stream segmentation	Yes **(7)** (7)	No **(3)** (3)	--
V6	K_h – UGA	High **(6)** (0)	Medium **(3)** (0)	Low **(1)** (10)
V7	Top surface of the PSU	Uniform surface **(5)** (7)	Interpolated surface **(5)** (3)	--

Table 4.6.: Variable features and their states (bold numbers are the count of top 10 models containing the given feature state; regular numbers are the count of bottom 10 models)

The table indicates that the model configurations of the top and the bottom models had distinctive configurations and certain variable feature-states were more/less prevalent in these configurations.

The preference was the complete for the following feature-states:

- All top 10 models contained state V12 (variable thickness of the bottom of L1),
- All bottom 10 models contained state V11 (uniform thickness of the bottom of L1),
- All bottom 10 models contained state V32 (3-zone configuration of the PSU),
- All bottom 10 models contained state V63 (low permeability set for the UGA).

The preference was the decisive for the following feature-states:

- Segmented streams (V51) were the preferred configuration in 7 of the top as well as the bottom 10 models, while an aggregate stream (V52) was preferred in the remainder of 3 models in each model set.
- Uniform top surface for the PSU (V71) was found in 7 out of the bottom 10 models, while the remainder of the 3 models contained the variable top surface feature state (V72).
- The high permeability set (V61) was more preferred in the top 10 models than the other two feature states.

There was near to total equivalence for the following feature-states:

- Both states of bottom of layer 1 (V1) featured almost equally in the bottom 10 models.
- Both states of bottom of layer 2 (V2) featured almost equally in the top 10 models.
- Both states of the extent of the PSU (V3) featured almost equally in the top 10 models.
- The no recharge condition (V53) was preferred in 5 of the top 10 models.

Different variable states were used to configure multiple models. Boxplot and ANOVA analyses were carried out to measure how the A* values changed with respect to different states of different variable features and to investigate which states brought significant changes in the A* values.

The A* values of the 200 models were divided into groups of the states of each of the 7 epistemic variable features. Firstly, boxplots were generated from the A* values and visually compared to assess the differences among the states (Figure 4.14-a-g). The full range of V12 boxplot spread wider than that of the V11 boxplot, and also the median A* values noticeably varied (Figure 4.14-a). The interquartile range in V21 boxplot was larger than that of the V22 boxplot; the median A* values were slightly higher in the former (Figure 4.14-b). Boxplots for V31 and V32 showed similar ranges; larger median A* values were found in V32 than in V31 (Figure 4.14-c). Boxplots of three states of variable V4 (V41, V42, and V43) appeared similar

with respect to their spreads and their median A* values (Figure 4.14-d). Also, the boxplots of the two states of variable V5 (V51, V52) appeared similar with respect to their spreads and their median A* values (Figure 4.14-e). V63 boxplot showed the largest interquartile range, followed by V61, and then by V62. The median A* values decreased from V63 to V62 to V61 (Figure 4.14-f). The interquartile range in V71 boxplot was larger than that of the V72 boxplot, but the median A* value was higher in the later (Figure 4.14-g).

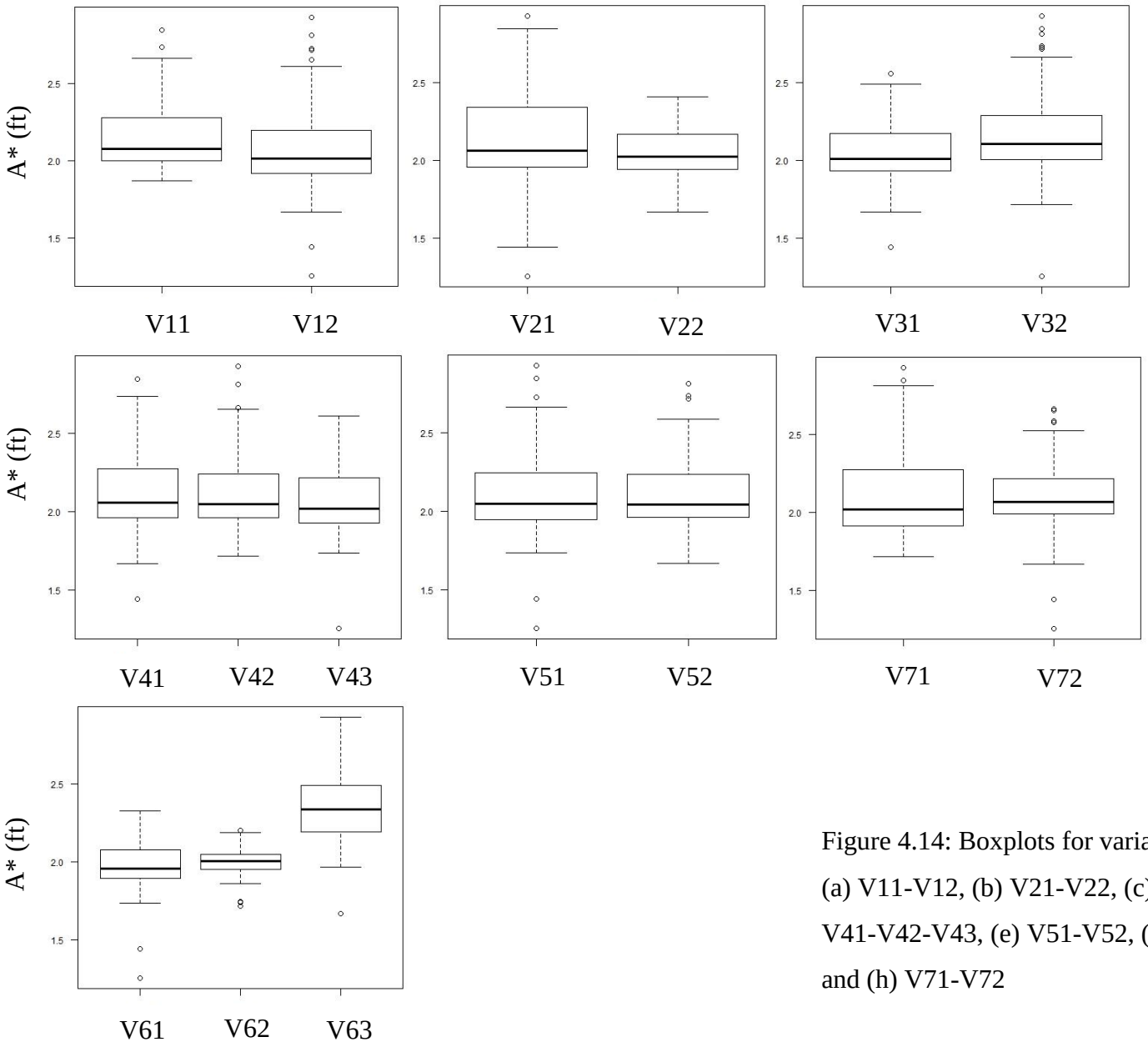


Figure 4.14: Boxplots for variable feature states (a) V11-V12, (b) V21-V22, (c) V31-V32, (d) V41-V42-V43, (e) V51-V52, (e) V61-V62-V63, and (h) V71-V72

The statistical significance of the above differences in A* values with respect to different state group was evaluated by one factor unbalanced ANOVA test (Table 4.7). The test was unbalanced because of the unequal number of A* values encompassed in most of the groups.

The difference in the A* values between V11-V12, V21-V22, V31-V32, and V61-V62-V63 was found to be statistically significant in the one-way unbalance ANOVA test. Both, the boxplot and the ANOVA analysis, indicated that feature-states "V12" (constant thickness of layer 1), "V21" (position of the bottom of layer 2 close to the bottom of layer 1), and "V32" (3-zone configuration of the potentially semi-confining unit) were individually part of the configurations of models with lower A* values. Also, feature state V61 – higher hydraulic conductivities for the 3 UGA layers – was found to be statistically significant. All these features highlight the importance of reasonable representation of the hydrogeologic units and their characteristics. For instance, changing the position of the bottom of a layer changes a layer thickness that, in turn, changes the transmissivity of the layer (and that of the adjacent layer) that is the product of the thickness of the aquifer its hydraulic conductivity. Additional information, such as more geologic borings could be used to configure the vertical discretizations of the UGA more accurately. Also, the heterogeneity in hydraulic conductivity could be further discretized using horizontal and vertical zoning of the conductivity values.

V11V12	V11 (n=93), V12(n=107)				
	Df	Sum.Sq.	Mean Sq.	F value	Pr(>F)
Feature	1	0.2769	0.276853	4.6763	0.03178*
Residuals	198	11.7222	0.059203		

(a)

V21V22	V21 (n=117), V22 (n=83)				
	Df	Sum.Sq.	Mean Sq.	F value	Pr(>F)
Feature	1	0.5284	0.52844	9.1213	0.002859**
Residuals	198	11.4706	0.05793		

(b)

V31V32	V31 (n=101), V32(n=99)				
	Df	Sum.Sq.	Mean Sq.	F value	Pr(>F)
Feature	1	0.570	0.57003	9.8755	0.001932**
Residuals	198	11.429	0.05772		

(c)

V41V42V43	V41 (n=76), V42(n=72), V43(n=52)				
	Df	Sum.Sq.	Mean Sq.	F value	Pr(>F)
Feature	2	0.1986	0.099308	1.6579	0.1932
Residuals	197	11.8004	0.059901		

(d)

V51V52	V51 (n=94), V52 (n=106)				
	Df	Sum.Sq.	Mean Sq.	F value	Pr(>F)
Feature	1	0.0001	0.000069	0.0011	0.9731
Residuals	198	11.9990	0.060601		

(e)

V61V62V63	V61 (n=55), V62(n=76), V63(n=69)				
	Df	Sum.Sq.	Mean Sq.	F value	Pr(>F)
Feature	2	6.0800	3.04000	101.18	2.2e-16***
Residuals	197	5.9191	0.03005		

(f)

V71V72	V71 (n=91), V72 (n=109)				
	Df	Sum.Sq.	Mean Sq.	F value	Pr(>F)
Feature	1	0.0021	0.002059	0.034	0.8539
Residuals	198	11.9970	0.060591		

(g)

Table 4.7: ANOVA results for (a) V11V12, (b) V21V22, (c) V31V32, (d) V41V42V43, (e) V51V52, (f) V61V62V63, and (f) V71V72 states (*= p<0.5, **=p<0.01, ***=p<0.0001)

The differences in the A* values for other variable features: V4 (configuration of the recharge basins), V5 (stream segmentation), and V7 (surface interpolation of the PSU), were not statistically significant. The impact of the change in the recharge basins' configuration, ("V4") would be limited in the areas surrounding the southern perimeter of the landfill property and not observed further downstream or upstream areas of the model.

Variable feature "V7" represented the uncertainty related to the PSU configuration; the effects of this uncertainty, if any, were represented by one well, S96202. This well was screened in the region where the transition between the Upper Glacial aquifer to the PSU, and from the PSU to the Magothy aquifer takes place. A detailed inspection of the A value descriptors indicated that the geology at this location was perhaps misrepresented and therefore the simulated heads at this well do not match well with the observed heads at multiple system states. However, a singular well may have underrepresented these differences between the states. For example, the A* values were recalculated excluding well S96202 (using 132 wells) and ranked using the new A* values with respect to the 200 models. Table 4.8 shows the top 10 models and their original and revised ranks. The difference between the original A* values (included S96202) and new A* values (excluded S96202) was not noticeable. Also, the difference between the original model ranks and the new model ranks was not noticeable.

Model #	Original A* (feet)	New A* (feet)	Original Rank	New rank	Original A* - New A* (feet)	Original rank - New rank
178	1.257	1.216	1	1	0.041	0
265	1.441	1.420	2	2	0.021	0
200	1.667	1.623	3	3	0.037	0
244	1.717	1.725	4	6	-0.007	-2
177	1.738	1.739	5	7	-0.002	-2
204	1.740	1.705	6	4	0.035	2
216	1.745	1.709	7	5	0.035	2
242	1.814	1.821	8	8	-0.006	0
213	1.827	1.829	9	10	-0.002	-1
141	1.830	1.833	10	13	-0.003	-3

Table 4.8: Change in A* and model ranks for the top 10 models with respect to inclusion/exclusion of well S96202

The impact of variation in stream segmentation, shown by two states of variable feature V5, was not found to be statistically significant. The streams in the model domain are gaining streams, that is, the flow of water into these streams is dependent of the changes in the surrounding groundwater levels (Peterson 1987). The streams could be affected by the fluctuations in the water table, but the reverse may not be true. Therefore, the impact of stream segmentation on groundwater heads, and subsequently on the area metric values, could be negligible and not detected in the ANOVA test. The impact of stream segmentation could be accounted if the area metric values are calculated with respect to different system response quantity such as the streamflow volume that is a better representation of the stream dynamics.

4.6. Association between A* and RMSE

As mentioned earlier, the common method used in groundwater modeling to measure models' representativeness or their replicative validity is RMSE. RMSE is calculated for a singular observation data set and therefore the RMSE values calculated for the models could not be directly compared with the overall area metric values calculated for these models. Therefore, three RMSE values were calculated for each model for each water table condition (high, median, and low); these RMSE values were then averaged to arrive at a singular, average RMSE value for each model.

Figure 4.15 shows the scatter plot where the models are potted on the basis of their average RMSE score with respect to their corresponding overall area metric values. A linear trendline was added to highlight a noticeable association between the two performance measures and the correlation coefficient indicated the strength of the correlation to be 0.657.

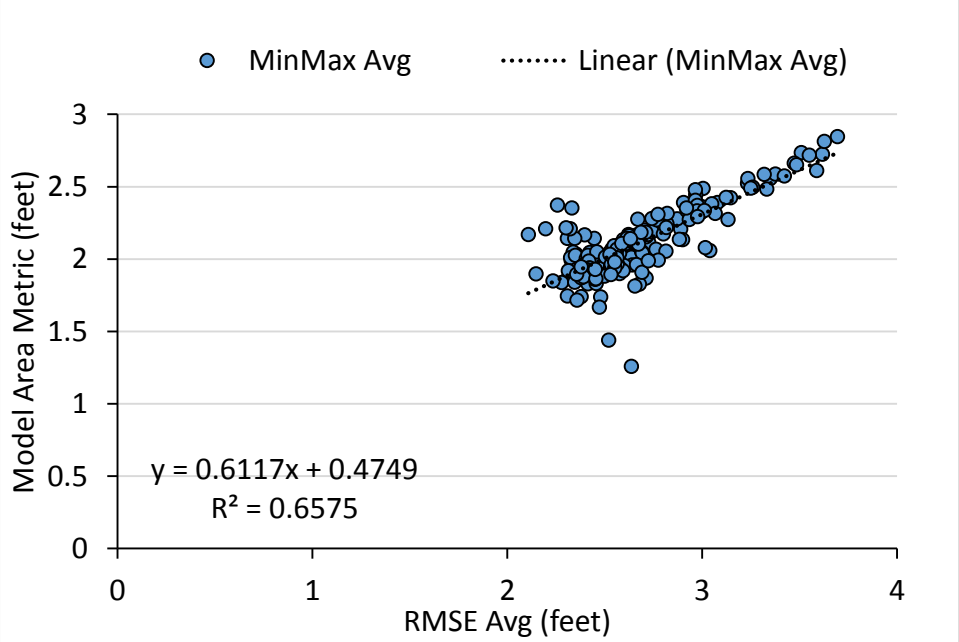


Figure 4.15: Scatter of the models' values of the model area metric (in feet) on the basis of models' average RMSE values (in feet)

In addition, the models were ranked separately on the basis of both, their overall area metric values as well as their average RMSE scores. Figure 4.16 shows the scatter of the 200 models on the basis of their corresponding overall area metric values with respect to their

average RMSE score. The figure indicates that the model rankings are comparatively closely associated towards the top right compared to the bottom left where there was more scattering. Certain model ranks changed significantly, for example, the model that was ranked number 1 with respect to the overall area metric value (model #178) was placed at rank 120 with respect to its average RMSE score. On the other hand, the model that was ranked number 173 with respect to the overall area metric value (model #243) was placed at rank 5 with respect to its average RMSE score. The correlation between the ranks, as indicated by the correlation (0.493), could be described as weak to mild.

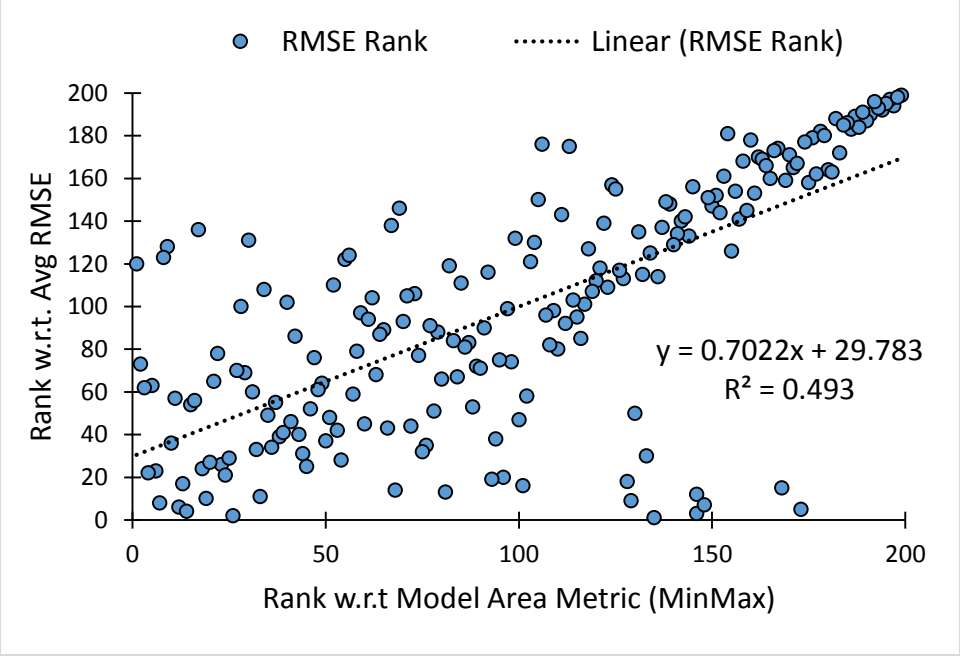


Figure 4.16: Scatter of the rankings of the models based on the average RMSE value (in feet) on the basis of model rankings with respect to the values of their model area metric (in feet)

The distinction between the traditional performance measures, such as RMSE, and the area metric is that the proposed approach acknowledges that the development of an exact model to represent the groundwater system is improbable, given uncertainty. Typically, the area metric is used for validation assessment of models that generate entirely probabilistic outputs that can be depicted as distributions (Ferson et al. 2008). Groundwater models are inherently deterministic models that are based on principles of conservation of mass, momentum, and energy (Konikow 1996). However, the explicit incorporation of uncertainty in the multiple

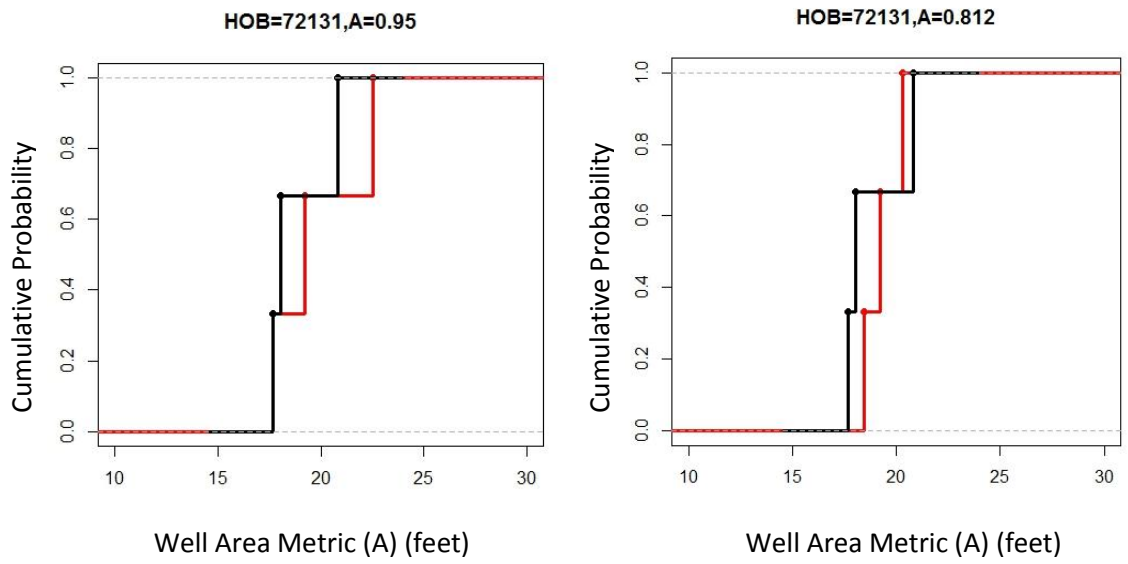
model approach and use of multiple sets of observed data creates distributions akin to the probabilistic outputs hitherto used in area metric studies.

The value of the area metric is dependent on the differences between the whole ECDFs of the observed and the simulated data. The model ECDF of a given model is a probabilistic representation derived from a set of deterministic values generated by simulating that model multiple times to represent different system states; these deterministic outputs for each state that are then then arranged into model ECDFs. Also, the observed data from different system states are collated into observed data ECDFs. The differences towards the tails of the distribution also affect the value of the area metric, not only the differences in the lower-order moments such as the mean and variances of true distributions. An ECDF accommodates the details about the system behavior that include "typical" behavior as well as the tails. Therefore, a distribution-based comparison facilitated by the area metric approach can be more informative than a comparison of means (Ferson et al. 2008). This avoids overfitting to a particular system state. Instead, it assesses the extent of a broader agreement with the collation of states.

4.7. Sensitivity Analysis

Two sensitivity analyses were conducted to assess the changes in the model performance with respect to changes in the descriptors and in their resolution (number of data points).

In the first experiment, the descriptors of the observed data that were used to calculate the A values were changed. The original descriptors of the observed data (the minimum, median, and maximum) were replaced with the quartile range: 1st quartile value, the median, and the 3rd quartile head observation value. The $ECDF_{observed}$ were derived from this quartile range for each well. The $ECDF_{simulated}$ remained unchanged. For example, Figure 4.17-a shows the $ECDF_{observed}$ generated using the original descriptors range, along with the $ECDF_{simulated}$ for well S72131. Figure 4.17-b shows $ECDF_{observed}$ generated using the quartile descriptors, along with the $ECDF_{simulated}$. The A value was 0.95 feet for the original descriptors, while the A value was 0.81 feet for the quartile descriptors for well S72131.



(a) (b)

Figure 4.17: $ECDF_{observed}$ generated using (a) the original descriptors, and that using (b) the quartile descriptors (in red), along with the $ECDF_{simulated}$ (in black) for well S72131

The A* values were then calculated for each of the 200 models as per the previous procedure. The A* values calculated using the original descriptors were subtracted from the quartile-based A* values of the same model (Figure 4.18). The quartile-based A* values were less than the A* values based on the original descriptors; the difference was < 0 for 179 out of 200 models. The differences ranged mostly from -0.8 feet to 0.2 feet, while the differences were between -0.2 to -0.75 feet in the top 7 ranked models.

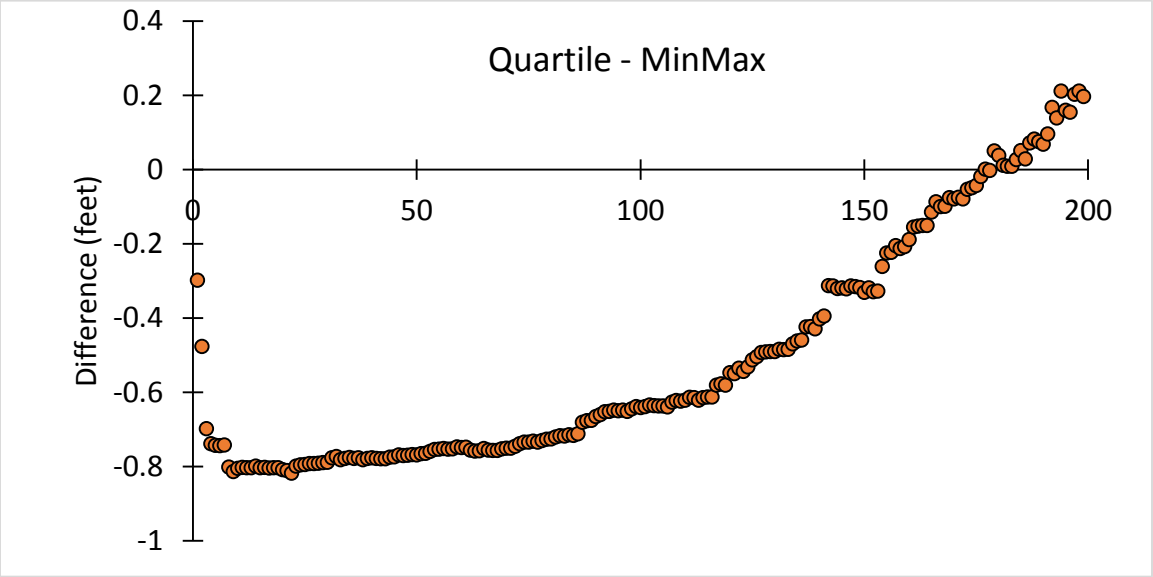


Figure 4.18: Difference between the A* values calculated using the original descriptors for a given model and the corresponding A* values calculated using the quartile descriptors

The range of the quartile descriptors was narrower than the original descriptors' range that included extreme values: the minimum and the maximum head observations. Models' A* values are a functions of the entire range of the ECDF. The comparatively narrower range of the quartile descriptors lowered the A* values compared to the A* values calculated with respect to the original descriptors. Models were ranked from smallest to largest with respect to the A* values in both cases. Figure 4.19 shows the scatter plot of the models' ranks based on the A* values calculated using quartile descriptors plotted against models' ranks based on the A* values calculated using original descriptors. The correlation between the two sets of ranks was strong ($R^2 = 0.69$). This was expected given the fact that the quartile ranges were nested within the

original range that was correspondingly conservative and encompassed the extremities in the head observations.

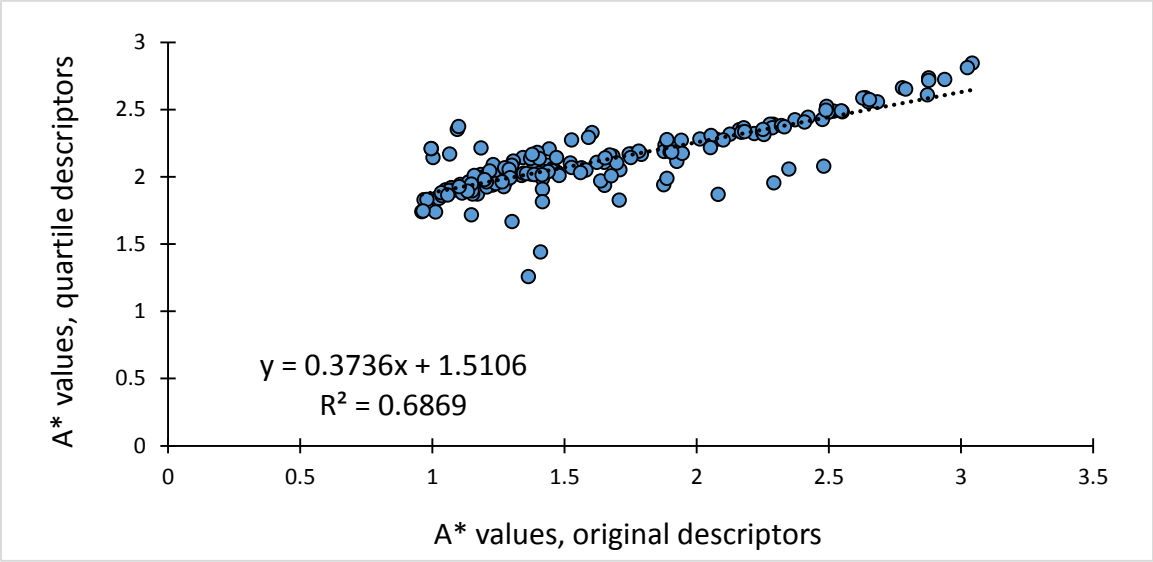


Figure 4.19: Scatter of the models' A* values (in feet) calculated using quartile descriptors plotted with respect to models' A* values (in feet) calculated using original descriptors

The second sensitivity analysis assessed the impact of change of the ECDF resolution from 3 vertical steps (3 data points) to five vertical steps (five data points). Only one model, model #178 that was ranked highest was re-simulated. Here, the three data points of the original descriptors – the minimum, median, and maximum head values – were replaced with five data points that included the minimum, 1st quartile, median, 3rd quartile, and the maximum head observation value for each of the 133 wells. The model was simulated for five (instead of the original three) groundwater levels by varying the value of the aleatory feature in the model – the northern constant head boundary (CHD1) five times: 42 feet (high groundwater levels), 41 feet (3rd quartile), 40 feet (median), 39 feet (1st quartile), and 38 feet (low groundwater levels). For example, Figure 4.20-a shows the $ECDF_{observed}$ and the $ECDF_{simulated}$ for well S72131 in the original, 3-step format. Figure 4.20-b shows the $ECDF_{observed}$ and the $ECDF_{simulated}$ using the five data points for the same well. Notice that the vertical cumulative probability of the ECDF was divided into five equal-length vertical steps. Five simulated values were generated that were collated into the $ECDF_{simulated}$ for each of the 133 wells. The A value for the wells were

calculated using the revised $ECDF_{observed}$. The A value was 0.95 feet for the original descriptors, while the A value was 0.81 feet for the quartile descriptors for well S72131.

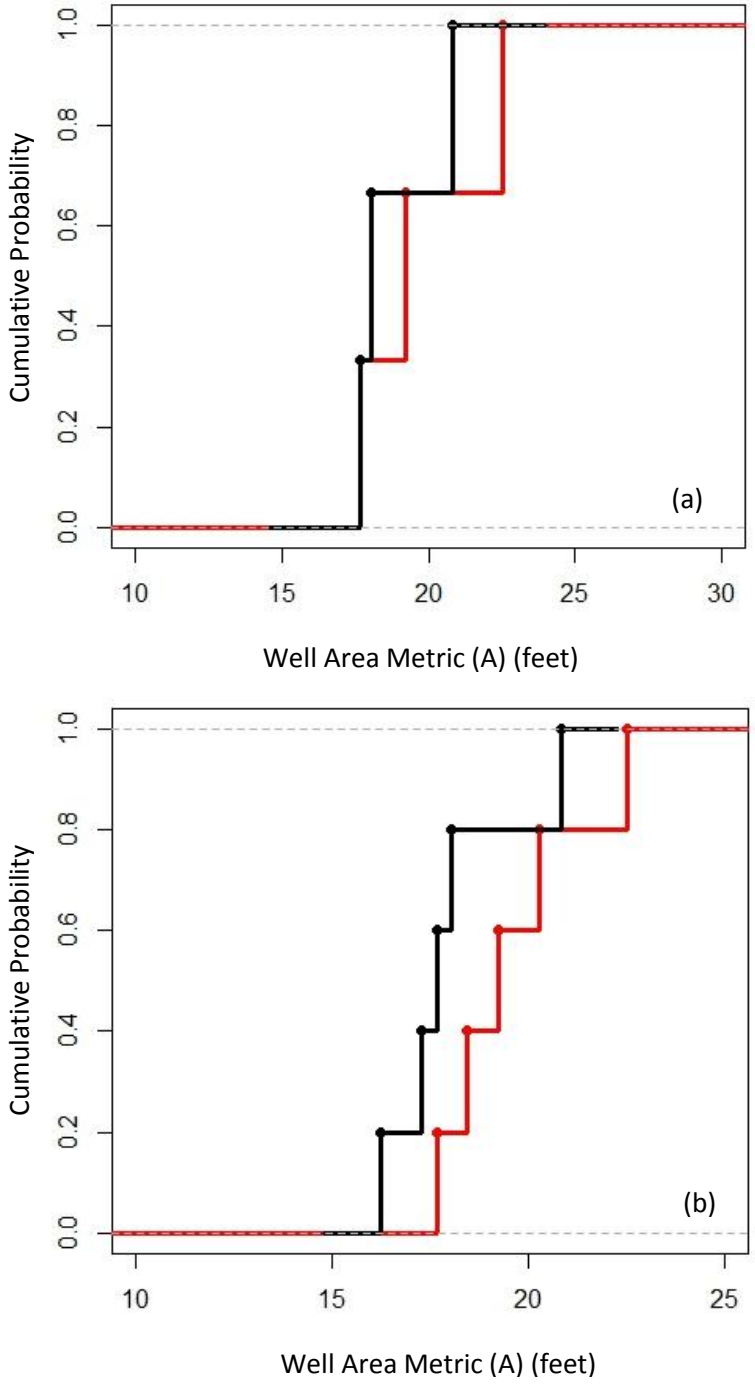


Figure 4.20: The $ECDF_{observed}$ (in red) and the $ECDF_{simulated}$ (in black) generated using the (a) original, 3 data points and, (b) new, 5 data points for well S72131.

The A values for the 133 wells were calculated using the revised ECDFs and compared with the corresponding A values calculated using the original ECDFs. Figure 4.21 shows the differences in the 5-step and the 3-step A values plotted in increasing order of magnitude for all 133 wells. The smallest difference was -1.087 feet for well S72164, while the largest difference was 3.52 feet for well S72167. The differences were net negative for 15 wells, and net positive for the rest of the wells. The difference was (<0.1 foot) for 19 wells and within $\pm$ 1 foot range 126 wells. The A* value for model # 178 increased from the original 1.25 feet to 1.69 feet in the revised procedure. It was concluded that the A values and the A* values were sensitive to the resolution of the $ECDF_{observed}$ and $ECDF_{simulated}$ and that higher resolution of the ECDFs may increase the As and the A*s.

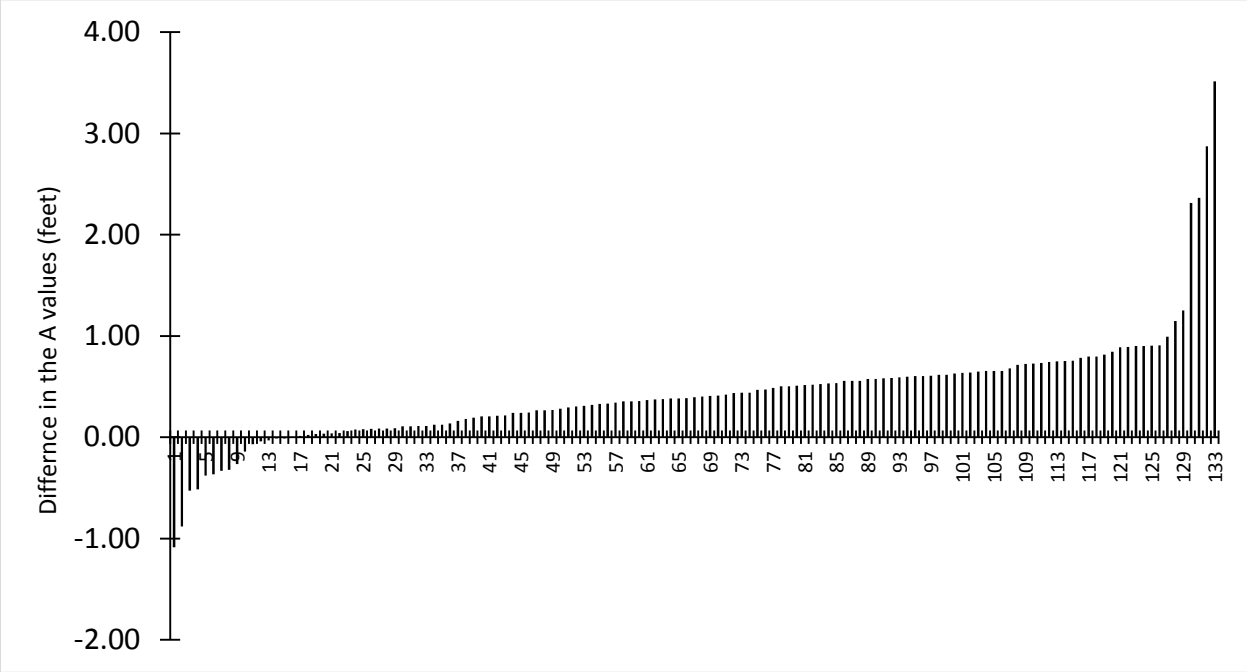


Figure 4.21: Differences in the five-step and the three-step A values for all 133 wells

Increasing the resolution of the ECDFs would increase the computational burden because each model variant would have to be simulated multiple times for each specific system state. Whether this increase in the computational burden is justified (or not) depends on the differences in the system's states. The increase in the resolution (from 3 steps to 5 steps and beyond) is justifiable if each system state is noticeably distinct from the other and therefore requires a separate model simulation. If the fluctuations of observed heads are marginal, say, $\pm$ 2/100th of a foot, then perhaps a singular model representation may suffice. Increase in the steps in an ECDF increases the resolution of the aleatory uncertainty associated with the system. Each additional step would bring the observed data ECDF closer to incorporating more of the inherent fluctuations of the system. However, each additional step would also require more accurate data so that these multiple system states can be accurately and distinctly represented. Additional data requirements increase the resource intensity of the modeling project. Therefore, decision should be made with regard to the relative worth of incorporating additional aleatory uncertainty. One possible solution could be to vary the resolution using trial and error until the A values stabilize. Such a decision could depend on the objective of the exercise. Here, the decision to restrict the resolution to three steps was taken on practical grounds given the unequal number of data points associated with each observation well (ranging from 2 to 446).

Chapter 5

General Discussion

Overview

This chapter discusses broader implications and the generalizability of the area metric-based multi-model validation assessment ("the proposed approach"). The discussion is centered on the following salient features of the proposed approach.

1. The proposed approach allows generating and testing multiple models. This is better than using a fixed, singular model given model uncertainty.
2. The proposed approach estimates the degree of model's representativeness rather than resorting to the conventional hypothesis testing approach, where models are either validated or invalidated, that may be unachievable given the model uncertainty.
3. Model uncertainty may not be reduced or eliminated. Therefore, the proposed approach adopts a more pragmatic approach by recognizing model uncertainty, incorporating it into the models, and then evaluating its impact on model representativeness.
4. The proposed approach allows a blind assessment of the models' representativeness instead of adjustment or updating them to fit a particular observational data set. This prevents over-tuning or over-fitting of the model to the reality.
5. The proposed approach allows validation assessment over a range of observed data rather than testing the models' performance over a singular observational data set that may be a snapshot representation of reality. This range represents different systems states and is depicted as an empirical cumulative distribution function (ECDF).
6. The area metric is flexible; it is non-parametric, mathematically well-behaved, and independent of the number of data points. It is easier to understand because it can be represented graphically and in the same units as that of the observed data.
7. The proposed approach can be applied for validation assessment of models where more than one and different types of observed data are available. It can also be applied in transient state models for pattern matching. The modeling solution obtained for the inverse problem from using the proposed approach can be used to solve the forward problem, such as prediction of contaminant fate and transport in case of the landfill model.

In the previous chapter, the multi-model validation assessment using the area metric approach was used in the context of simulating the flow of groundwater through the unconsolidated deposits surrounding the Brookhaven landfill in south-central Long Island. This chapter discusses broader theoretical and methodological implications, and the generalizability of the area metric-based multi-model validation assessment ("the proposed approach") to other, similar applications of simulation modeling. The discussion highlights seven salient features of the proposed approach and expounds on how the proposed approach is better suited for a multi-model validation assessment on the basis of these features. This elucidation is translated into recommendations on modeling practices for modelers engaged in similar modeling exercises elsewhere and who want to address model uncertainty in a manner that augments model users' confidence on the modeling exercise while maintaining its scientific objectivity.

5.1. Using and Testing Multiple Models

Developing a singular model is problematic for three main reasons. First, the data that are used to configure such a model may not necessarily collected for the purpose of model conceptualizations and development. The historical nature of parts of the data collection, and the fact that the data recording programs may follow different protocols for data collection, would mean that there would be no opportunity for to ensure data quality or to investigate measurement errors. These data may not represent all states of the hydrogeologic system, but only a limited range of values. Incomplete, infrequent, missing, and potentially erroneous data make it difficult to identify the unique or the perfect model from the model space, if such a model exists. Second, subjectivity is inherent to modeling and it is incorporated at the conceptualization phase (Walker et al. 2003) because models are a function of the modeler's world view (Barlas and Carpenter 1990). Different modelers can conceptualize groundwater system differently and develop models that reflect their subjective interpretation of reality. Third, using a singular model makes no room for incorporating the model uncertainties and hence makes no scope for testing and differential understanding of the system. Although the immutability of the singular model is assumed, these models are often calibrated or updated to achieve a better numerical fit with the observed data. For these reasons, it is difficult to justify a particular single model as the unique modeling solution.

On the other hand, developing and testing multiple models allows us to test our theories and our understanding of the structure and functioning of the system being modeled. As mentioned in Chapter 1, developing multiple model variants has resonates with Thomas Chamberlin's "multiple working hypothesis" where he cautioned against developing a preference for a particular hypothesis (read, model) and recommended that multiple hypotheses should be developed to achieve completeness in the investigation (Chamberlin 1965). In his words:

> "*But in a single working hypothesis may lead investigation along a given line to the neglect of others equally important; and thus, while inquiry is promoted in certain quarters, the investigation lacks in completeness. But if all rational hypotheses relating to a subject are worked co-equally, thoroughness is the presumptive result, in the very nature of the case."*

Similarly, multiple interpretations of the groundwater flow regime were translated into multiple models. These models were considered equally plausible initially, and then subsequently ranked on the basis of their degree of validity. Given a range of models were developed and tested, the Brookhaven landfill groundwater flow simulation modeling exercise is more complete than developing a singular landfill model and then adjusting/updating it to achieve greater numerical fit. The model variants evaluated here represent a cluster of models from a nearly infinitely large model space of unique models, each one representing reality with greater or lesser degree of overlap. This sample may not be exhaustive and unbiased (Rojas et al. 2008). However, increasing the space of potential candidate models also increases the chances of bracketing the model that is the closest estimator of the true model, given the model uncertainty. Also, validation assessment using the proposed approach highlight which, among the space of potential model solution, are better representations and which are among the worse. Further exploration could reveal the reasons for models' variable performances, such as the difference in their configurations.

Therefore, a multi-model analysis, rather than vouching for the truthfulness of a singular model, is recommended as a more preferable choice to the modeler elsewhere who face the similar challenge of developing a more representative model under uncertainty.

5.2. Assessing the Degree of Model Validity

Several definitions of the term model "validation" are used and a diversity of opinions persists about what is meant by it and its intended use (Sargent 2009; Ferson et al. 2008; Oberkampf and Barone 2006; Law 2005; Law and Kelton 2000, p. 264; Balci 1998; Barlas 1996; Refsgaard and Knudsen 1996; Forrester and Senge 1980; Schellenberger 1974). The preceding question – "Can models be validated?" – has been widely debated (Oreskes et al. 2010, 1994; Sterman 2006; Bredehoeft 2005, 2003; Oreskes 2003, 1998; Barlas 1996; Tsang 1991; Barlas and Carpenter 1990; Forrester 1961). Two schools of thought have emerged – "verificationism" and "falsificationism".

Verificationism leads to claims that "the model has been validated". Models are either "accepted" or "rejected" (Barlas and Carpenter 1990). Under verificationism, if a model is the simulator of reality, then the output will match real-world data; conversely, if output matches observations, then the model is the simulator of the reality. If the model error is found to be insignificant, then the model is said to be validated. Falsificationism states that models cannot be validated, only invalidated (Popper 1959, p.27). Under falsificationism a model is tested, and conceptual error or incongruence of output and observations result in its rejection as an acceptable simulator of reality. Falsification rejects models for not reasonably simulating reality, but it does not "accept" models even if the simulated outputs overlap (Oreskes et al. 1994; Konikow and Bredehoeft 1992). Confidence regarding the model's validity increases as more tests fail to reject the model. However, no amount of confidence is tantamount to endorsement ("validation") of the model.

Philosophically it may be that theories, laws, and models can only be invalidated and not validated, disproved and not approved, and falsified and not verified. However, it has been argued that such a perspective might be "unproductive or even debilitating" for assessing the representativeness of models in the fields of engineering and natural sciences (Oberkampf and Roy 2010, p. 401). The area metric is a descriptive measure of the model's operational replicative validity; that is, it quantifies the difference between observed and simulated data, such that smaller values of the area metric mean smaller differences. Its values remain unchanged if the observed and simulated data remain the same, regardless of which analyst calculates the metric (Ferson et al. 2008). However, the adequacy of model's replicative or

predictive validity is not decided by the area metric. In other words, it does not determine if the model performance is good or not. It is recommended that a modeler applying the proposed approach should remember that this decision is the purview of the decision maker. So, Wexler and Maus (1988) judged the USGS plume model subjectively: "*Comparison of the simulated with the observed water-table altitudes show that the water levels match closely over the entire study area*." Likewise, a determination of whether an area metric score is "satisfactory" is left to model users' judgement.

It is important to note that the area metric is for "accuracy assessment" and not for "adequacy assessment" (Oberkampf and Barone 2006). The area metric based multi-model validation assessment can be deemed as "scientific validation", or quantification of the models' accuracy independent of the modeling objectives or the needs of the modeling project. This is different from "project-oriented validation": that is, the accuracy quantification undertaken with the modeling objective in mind, and that is contextualized with respect to the needs of the modeling project (Oberkampf and Roy 2010, p. 374). This separation is necessary to maintain the objectivity of the area metric (Ferson et al. 2008). The area metric serves best as a support tool for the model user's decision, in accordance with the requirements of the modeling exercise.

Second, the binary validation or invalidation of a model requires complete and deterministic understanding and no uncertainty about the system that is being modeled. However, such a situation is extremely rare and unachievable in case of a real world model exercise. Here, the proposed approach accounts for heterogeneities, uncertainty, and the scale of data. Specifically in case of the landfill model, a singular, deterministic model could not be built from the observed data given uncertainty in these data. Uncertainty in the hydrogeologic settings was acknowledged and explicitly incorporated in the solution space. Hence, 288 model variants were generated and simulated; these model variants represented the solution space to the inverse problem where each model variant was treated as a potential solution. Also, the degree of the validity of these solutions was assessed over a range of observed data representing different states of the groundwater system, acknowledging that a singular, snapshot observed data set may not represent the variety of states assumed by the groundwater system. In addition, the evaluation of a given model is conditional on the data that are used in the validation assessment. Any change in the observed data would lead to a different assessment of model's validity. Therefore,

the data that were used to distinguish among several different models themselves may not be adequate for this identification purposes (Beven 2012).

Therefore, the proposed approach improves upon the conventional hypothesis testing that forms the basis for the verificationist and falsificationist approaches. Here, it is acknowledged that complete and deterministic understanding of the objective reality is infeasible and consequently the binary judgement of model's validity ("the model is valid" or "the model is invalid") is impractical. Therefore, unlike conventional hypothesis testing, where a model is either valid or not valid (invalid), the proposed approach evaluates a degree of model validity, as quantified by the level of agreement between the observed and the simulated data, represented by the distribution functions. Models that obtain smaller area metric score are ranked higher indicating better correspondence to observed data relative to models that have larger area metric scores and thereby are ranked lower.

This model users' evaluation of model's adequacy entails two types of risks: Type I risk (model builder's risk of rejecting a valid model) and Type II risk (model user's risk of failing to reject an invalid model). In practice, the likelihood of model user's risk is more than that of the model builder's risk.

First, end users of models are often unfamiliar with the process of model development, and the capabilities of what the models can and cannot do (Brugnach et al. 2007). Simulation modeling is a specialized activity. Translating the modeling exercise from the technical jargon to lucid colloquial language with full disclosure is challenging. One danger in such imprecise communication is that the nuances surrounding the term validation may be lost. According to Konikow and Bredehoeft (1993, p. 178):

> "*To the general public, proclaiming that a groundwater model is valid carries with it an aura of correctness that we do not believe many of us who model would claim. We can place all the caveats we wish, but the public has its own understanding of what the word implies. Using the word valid with respect to models misleads the public; "verified" carries with it similar connotations as far as the public is concerned*."

In addition, identification of model validity creates an aura of authenticity that can sideline further critical evaluation of the model. The end users of models are often unfamiliar

with the process of model development, and the capabilities of what the models can and cannot do (Brugnach et al. 2007). Informing the decision maker post-facto of the uncertainties leaves the decision maker in a position where significant money, time, and effort has been spend thinking of they will have "an answer" as the end product of the modeling exercise. The danger in such imprecise communication is that the nuances surrounding the term validation may be lost. Also, more evaluation requires expenditure of more time and resources, usually not a welcome proposition for administrators, managers, and funding agencies. Co-workers and management often do not welcome challenges to confirmatory biases, especially in competitive, commercial environments.

To minimize model user risk, it is recommended a modeler should be familiarize the stakeholders with the modeling process, and they should be periodically updated about the direction and progress of the modeling exercise. Also, all documentation, modeling protocols, and modeling data should be made available (Morel-Seytoux 2001). Simulation modeling is a specialized activity. Translating the modeling exercise from the technical jargon to lucid colloquial language with full disclosure is challenging. Therefore, regular engagement with model users can minimize the element of surprise that may result at end-of-the-process evaluation of the model, when considerable resources have already been spent.

The second contributor to increase model user's risk is that many model users are not experts in hydrogeology and they rely on the modeler's self-certification of the validity of the model. This raises concerns regarding the "critical distance" that a modeler maintains with the model; this critical distance is the ability of the modeler to impartially examine the uncertainty and the indeterminate elements in models. Excessive professional and emotional involvement of modelers with their models reduces critical distance. This involvement stems from spending considerable time and resources in developing the model, and the practical issue that the performance of the model relates to funding. Good modeling practices should be followed to generate models that give right answers for the right reasons. For example, when a desired outcome is forced at a grid node by manipulating boundary conditions, a good match does not indicate that the model is representative, in fact, such a match has little meaning or value (Konikow 1996). Simulating complex real world systems is a challenging task that may involve the modeler emotionally, so that the persuasive power of the model means the modeler may start

to “believe” in the model, losing the capability for dispassionate, critical examinations of model’s soundness. Reduced critical distance may lead to the modelers trying to present their work most favorably, becoming strategic in presentations of the model and its associated documentation, suppressing its shortcomings, and generally overselling the product (Lahsen 2005).

Certain precautions are recommended for a modeler to minimize risks posed by reducing the critical distance. For instance, models should be peer-reviewed, in a transparent and constructive manner, by independent reviewer. One way to assess the representativeness of groundwater models is via “post-audits”. Post-audits compare the predicted output with observations after a given time, when the corresponding system response may actually have been manifested (Konikow 1996). Very extensive history matching is another commonly employed technique. For instance, in this study the groundwater head observations made at wells were compared over 30 year time horizons. However, note that errors in assumptions and interpretations are not eliminated if the validation is based on data of inconsistent quality, no matter how large or extensive the data sets may be. No amount of validation assessment can guarantee selection of the best possible model; however, additional experimentation or failure of concurrence between the model and future data may expose model inadequacies. Incorporating additional information regarding transient hydrologic conditions could constrain the uncertainty associated with the model space and, in turn, the models’ predictive parameter estimates (Rojas et al. 2008; Freer et al. 1996). Therefore, it is recommended that modeler using the proposed approach should err on the side of the caution.

5.3. Incorporating Model Uncertainties

The primary goal of incorporating these uncertainties was to test their impact on the models' performances, as reflected by the model area metric (A*) values. As mentioned earlier, the uncertainty associated with eight model features was acknowledged and incorporated into the modeling exercise. Seven of these features were classified as epistemic variable features while one feature was classified as the aleatory variable feature. The uncertainties were represented by two to three states for each variable feature. The degree of these models' validity was tested using the proposed approach. These models incorporated the following model uncertainties: (i) uncertainty in geological structures (e.g. presence and extent of the PSU); (ii) model parameters that characterize the variation of the hydraulic properties (e.g. three sets of K values); and (iii) local heterogeneities of model parameters (e.g. variation in the liner configuration). Additionally, boundary conditions (e.g. northern CHD boundary) were altered to see impacts on the model-simulated output. Other features were fixed. Multiple models variants were generated using a combinatorial approach where different states of epistemic and aleatory variable features were combined, along with the fixed features.

Uncertainty in select model features is acknowledged and then incorporated in the form of multiple models of varying compositions. Here, eight variable features were introduced into the model. The acknowledgment of uncertainty in select features was dependent on the preliminary assessment of the influence of these features during the pilot iterations, as well as through consultations with subject matter and site specific experts. Other model features were kept fixed to limit the scope of the modeling exercise. There can be several other model features of the model domain that may be fully or partially uncertain, or have been interpreted differently by different modelers. Each interpretation can be translated into a working model of the groundwater flow system and can be subsequently evaluated. The number of model conceptualizations that can be assessed increases as uncertainty in additional model features is recognized and included in the model. Consequently, model rankings could change with further inclusion, or exclusion, of unacknowledged model uncertainty (Oberkampf et al. 2002). Whether acknowledged, or not, uncertainty in a model feature is a subjective decision, and could vary if the purpose and scope of the model were to be changed.

In addition, the acknowledged model uncertainties were classified as reducible (epistemic) and non-reducible (aleatory) uncertainties. This classification approach is more practical compared to classifying the uncertainties in terms of their location or their magnitude. Location-based classification (conceptual, input, and parameter uncertainties) is problematic because the difficulties in apportioning model error to its constituent components (input error, parameterization error, and conceptual error) (Morel-Seytoux 2001) given variable features included in the model may interact with each other and may either compensate or confound model errors. On the other hand, the magnitude-based uncertainty classification requires a knowledge of how severe the uncertainty is on the scale ranging from complete deterministic knowledge to total ignorance. The magnitude of uncertainty in the variable features included in the modeling exercise was assumed to be equal for all variable features. The model area metric (A*) value is an aggregate quantitative evaluation of the accuracy of the computational models with respect to various modeling assumptions and approximations entailed in the computational model (Oberkampf and Barone 2006). Therefore, the A* score of a model was not fractioned and attributed to individual location of uncertainties. However, the comparative analysis of the composition of the top 10 and the bottom 10 models, as well as the boxplot and the ANOVA analyses of the variable features' states indicated that certain model feature-states that defined the local heterogeneities in the models' hydrogeologic settings and aquifer characteristics (V12- uniform thickness of layer 1, V21- position of the bottom of layer 2 close to the bottom of layer 1, V32-3-zone configuration of the PSU, and V61- high preamble configuration of the UGA conductivity) played an important role in lowering the A* values. Given that all these variable features are epistemic (reducible) uncertainties, the future data gathering efforts could be focused on improving our understanding about the local hydrogeologic heterogeneities and aquifer characteristics specified above.

Therefore, a modeler using the proposed approach would acknowledge that model uncertainty may not be reduced or eliminated. Instead, the modeler would adopt a more practical approach that involves subjective recognition of model uncertainty, its incorporation into the modeling exercise in the form of multiple model variants, and then the subsequent evaluation of the impact of model uncertainty on model representativeness by assessing the degree of models' validity using the area metric.

5.4. Blind Assessment

Typically, model calibration exclusively emphasizes parameter uncertainty. It is assumed that the model conceptualization and the calibration data are error-free (Jansen 2003). The model calibration process results select the best parameter set with respect to the performance measure used in the calibration program (Wagener and Gupta 2005). State-of-the-art automated calibration procedures, such as PEST (Parameter Estimation) (Doherty and Hunt 2010) or UCODE-2005 (Poeter et al. 2005), are commonly used in groundwater studies. The most common form of model calibration is localized factor perturbation where a particular parameter value is changed, keeping other values constant (Brugnach et al. 2007). Model calibration is iterative process of estimation and adjustment of parameters to improve the agreement between the simulated outputs of the model and corresponding observational data based on a pre-determined critical criterion at a chosen starting point (Rykiel 1996). The calibration continues until the difference between the two values meet pre-specified critical criteria. After each iteration, simulated values are compared with observed values. The results of the iterations are exposed to the observed data in order to make this comparison. If the difference between the two satisfied a certain critical criterion then the iterative process is terminated. A model is considered more representative as the differences approach zero.

Model calibration may compensate for, and thereby underestimate, the model conceptual uncertainty by optimizing the models' goodness-of-fit to the observational data that may not be error-free (Refsgaard et al. 2006, Reilly and Harbaugh 2004). Such over-tuned models may result in unreasonable predictions, irrespective of the goodness-of-fit at the calibration stage (Bredehoeft 2003). Improving the goodness-of-fit via standard as well as non-standard means may deviate of model conceptualization from the real world system being modeled (Reilly and Harbaugh 2004). An optimized model using a performance measure, such as the RMSE, may result in small variance between one set of model outputs and observed data. This fit may come at the expense of significant bias in the model (Moriasi et al. 2007). Trying to achieve a better model fit using standard as well as non-standard means may result in deviation of model conceptualization from the real-world system (Reilly and Harbaugh 2004). A well-fitted model (via calibration) that is conceptually flawed can make unreasonable predictions (Bredehoeft 2003). Over-fitting the model to observational data is also intended to compensate for conceptual

uncertainty (Refsgaard et al. 2006; Reilly and Harbaugh 2004; Bredehoeft 2003). Close matches to observational data can be achieved by more than one model by adjusting parameter values only, and not by rectifying conceptual flaws (Bredehoeft 2005). Typically, these parameter adjustments are annotated in model documentation, but sometimes they may not be so explicitly stated or explained adequately, intentionally or otherwise.

Also, improvement of the computational power is not a replacement for modeler knowledge of numerical methods and the hydrogeology of the study area. In the present study, the models were simulated using Visual-MODFLOW v.4.2 (Waterloo Hydrogeologic, Inc., 2006), a graphical user interface (GUI) for MODFLOW-2000. GUIs are an essential feature present-day hydrogeologic modeling because they make modeling user-friendly. However, they may increase the distance between the modeler and the core numerical method of modeling. Efforts to model subsurface processes, such as the groundwater flow, are handicapped because of lack of information, and this problem cannot be resolved though the computational power of simulation modeling has increased since MODFLOW was introduced in mid-1980s (Winter and Tartakovsky 2008). Lack of knowledge can introduce significant errors in the model solutions, and worse, these errors can go undetected because of the masking effect of typical model calibration.

Similar to model adjustment using calibration approaches, the prior distribution of the model input parameters can be updated using Bayesian analysis. Model updating may be a suitable and necessary objective in other cases of model applications but not in case of validation assessment. The primary goal of a Bayesian analysis is model updating that is different than the goal of evaluating the goodness-of-fit of multitude of models with fixed model configurations. Secondly, model updating may result in changes in the prior distribution of the model inputs while keeping the computational model fixed. Even so, an updated or posterior input probability distribution will be propagated through the same model that from which the prior distribution was propagated (Oberkampf and Trucano 2006). Assuming that the computational model is correct and thereby keeping it fixed would be equivalent of ignoring the conceptual model uncertainties. Undertaking Bayesian analysis is computationally more efficient when limited to updating the prior distributions of the parameterized inputs, such as hydraulic conductivity, than to update the distribution for the structural configuration of a model, such as its geologic

framework. In other words, it is easier to acknowledge the uncertainty only in the parameters and to assume a fixed structural framework than to acknowledge the uncertainty in the structural framework as well. However, achieving a closer model fit via parameter adjustment while excluding the structural uncertainty may overcompensate for a potentially erroneous conceptualization.

Validation assessment based on the area metric was different than model calibration or model updating. Fine-tuning the model via calibration undermines the primary goal of the model validation process: to assess the representativeness of the model with respect to the real-world system (Oberkampf and Roy 2010, p.374). In the proposed approach, there was no post-simulation exposure to the observed data and the subsequent adjustment in the model configuration. The area metric-based validation assessment was a blind assessment of the models' validation, that is, once the $ECDF_{simulated}$ values were generated then these values, and thereby the configuration of the models, were not subjected to any iterative change (Oberkampf and Barone 2006). A blind-assessment prevents post-simulation exposure to the observed data, and the subsequent adjustment in the model configuration. The 288 models' configurations remained fixed, and not revised (calibrated or adjusted), after the evaluation of their area metric values to achieve better model fit.

A calibration-induced improvement in the model is attributed to the automated calibration algorithms than to the improvement in the modeler's understanding of the functioning of the hydrogeologic system. Unlike calibration, a blind-assessment does not absolve the modeler of his duty to study the real-world system, to adhere to good modeling practices, and make an honest attempt to generate a conceptually correct model. Also, the application of the proposed approach not only allowed evaluation of multiple parameters within particular conceptual model frameworks, but also allowed evaluation of multiple structural model frameworks. Therefore, the blind assessment carried out using the proposed approach would put more emphasis of developing sounder model conceptualizations rather than attempting to achieve, as Voss (2011) puts it, "*detailed data fitting of highly parameterized models by merely pushing buttons and adjusting knobs on a computer program*".

5.5. Composite Assessment over a Range of Data

In the proposed approach, the degree of models' validity was assessed with respect to the system response quantity that is a measurable or observable response of the real world system; here, it was the groundwater heads. Groundwater head surveys were carried out daily, monthly, quarterly, and at annual frequencies over 30 years, at various locations. However, the measurements were not equally spaced and were distributed unequally across the three decades of observation period. Not all wells were surveyed each time. Data are missing at certain observation wells for certain periods. Not all wells were constructed in the same time period and not all wells that were constructed have been continued. Therefore, the counts in the observational data vary for each well. For example, four observations were made at well 72119, while 446 observations are recorded in case of well S3529. To maintain parity among the observation wells, the ECDFs were developed using three descriptors of groundwater head (median, maximum and minimum head observation values) derived from the available records for each of the 133 wells. The model variants were simulated for each of the corresponding groundwater conditions (high, median, and low groundwater conditions) to calculate the well area metric (A) values. The three descriptors used to develop the observed and the simulated data ECDFs cover a range over which the groundwater heads at a given well historically fluctuated.

In the proposed approach allowed validation assessment with respect to the structured summary of the observed data. Hence, the validation assessment was made over a conservative range of conditions using the proposed approach. The area metric values would be smaller only if both the location and the shape of the observed and the simulated data ECDFs are similar. In addition, the model area metric (A*) values were calculated by collating the well area metric (A) scores obtained for the 133 wells in a model-ECDF and comparing it with a hypothetical, reference model's ECDF for each of the 200 models. The 133 wells were distributed in different parts of the model domain and also were screened at different depths. Hence, the model area metric (A*) represented a composite picture of a model's performance across different parts of the model domain as well as over a range of observed data. In addition, it is intuitive for the model user(s) to judge the performance of the model in the same units as that of the observed data (that is, in feet) (Ferson et al. 2008). Visual comparisons of the ECDFs also aid this intuition (smaller areas between the observed and simulated data ECDFs mean better representativeness).

5.6. Utility of the Area Metric

Typically, a model's representativeness can be tested using various performance measures that may judge different or similar aspects of the model's performance. For example, bias measures model accuracy (how well does the model predicts with respect to the mean output value) while the standard deviation indicates model's precision (the amount of variation inherent in the model output). Least-squares performance measures such as the Root Mean Squared Error (RMSE) combine bias and standard deviation into one value, while the Index of Agreement is a dimensionless, standardized version of the MSE (Janssen and Heuberger 1995, Willmott 1981). The 2-sample Kolmogorov-Smirnov (2KS) test statistic measures the maximum absolute difference between the CDFs of the observed and the simulated data. Different objective functions may provide different parameter estimates; generally there is no correct choice of an objective function (Voss 2011). It is likely that a model that performs satisfactorily on one performance measure may not perform as well on another criterion. It is suggested that instead of relying on a singular objective function, the performance of the model should be used with regard to several criteria (Refsgaard and Knudsen 1996).

The area metric used in the proposed approach is also flexible approach in assessing the model performance. First, the area metric offered a non-parametric approach to quantify the replicative validity of the model over a range of observed data expressed in the same units as that of the data. No assumptions are necessary regarding the statistical nature of these samples. Second, the area metric is independent of the number of the simulation and experimental samples and the number of data points that go into the well-ECDF or the number of wells that the model-ECDF can be composed of is therefore not fixed but changeable. The observed and the simulated data, regardless of this count, can be collapsed into an ECDF allowing model assessment over a range of these data. It is important to note that the observed data and simulated data ECDF can match if the locations and the shape of these ECDFs match. Unequal number of data points or different locations of the ECDF would lead to area metric values greater than zero. Here, the number of data points were three in each of the well-ECDF and 133 for each of the model-ECDFs. This allows the area metric values to be zero in case of a perfect overlap in the locations of the ECDFs. It should be remembered that this kind of numerical accuracy does not guarantee that the model aptly represents all aspects of the real world system. Third, the area metric

approach used in this study integrated the absolute difference between the ECDFs of the observed and the simulated data (Ferson et al. 2008). The area metric offers improvement not only over the least-squares approaches given that it could collate and use observations from different periods, but also over the 2KS distributional comparison by integrating the differences across the distributions and not just quantifying the maximum difference. Fourth, if additional output response data are available, then the area metric can be calculated for them and then these metric values can be collectively used to decide the degree of model's validity. For example, in addition to simulated groundwater heads, groundwater simulations generate additional outputs such as estimates of streamflow rate and volume, or the direction and the rate of groundwater flow through different aquifer units. Incorporation of additional, diverse types of data into the model development process enables rejection of models that were previously considered to be representative of the real world system (Beven 2001). Likewise, an integrated surface water – groundwater flow simulation model can be considered more robust if it can adequately represent (match) more than one system response, such as head measurements and the rate and volume of flow into streams.

These features of the area metric make it more suitable to be used as a performance measure in situations where a modeler may not be able to fulfill all the requirements and assumptions of other performance measures given the constraints on the available data and given model uncertainty. A modeler should remember that the area metric should be considered as a validation measure with respect to the application domain for which the model has been developed. The area metric is suitable to estimate the degree of a model's replicative validity. Smaller values of the area metric increase confidence about the replicative validity of a model. Validation assessment pertaining to other aspects of the model, such as the justification of choices of parameters and variables and processes in the model, or the general evaluation of the model's conceptual adequacy needs to be based on a holistic assessment of the model, using different validation criteria including assessment of its replicative validity.

5.7. Generalizability of the Proposed Approach

Here, the multi-model validation assessment using the area metric approach was used in the context of simulating the flow of groundwater through the unconsolidated deposits surrounding a landfill in south-central Long Island. The applicability of the proposed approach can extend beyond the scope of the case study landfill model.

For example, predictive application of the area metric requires extrapolation of the area metric values generated for the replicative validation into the future. This extrapolation is meaningful if the future system states are similar to the existing or observed conditions (Ferson et al. 2008). Typically, such continuity is rare; future conditions may vary considerably from existing or historic conditions, and the model's time frame may not exactly overlap in the prediction time frame (usually the former is shorter than the latter) (Bredehoeft 2003). Therefore, theoretically the area metric cannot be computed for predictive models because no observational data are available to generate the $ECDF_{observed}$.

However, it is to be noted that the proposed approach is essentially an inverse problem where the representativeness of a model is evaluated on the basis of available data. Once the multiple models are ranked using the area metric score, then the top ranking model(s) can be used to determine the solution to a forward problem where the future behavior of a system can be determined using model solution(s) established by the proposed approach in the inverse problem. Hence, the inverse solution established by the proposed approach can be linked with the forward problem of finding a predictive solution going beyond the current observations.

This theoretical linkage between the inverse and the forward problem could be translated into tangible solutions in case of the landfill model. The model rankings represent the solutions to an inverse problem because the proposed approach is used to rank models, from a model space, on the basis of their fit to the given observational data. A particular solution, say, the top ranked model (model #169), can then be used as a base model to develop a simulation model that would either predict the fate and transport of the landfill leachate into the surroundings or to predict the response of the system to remedial measures taken to contain the leachate flow. The inverse solution, when vetted by the proposed approach, could improve the representativeness of the predictive model leading to an effective forward solution.

This blind assessment followed in the proposed approach could precede model calibration or updating as a screening mechanism. This screening process would include all model variants regardless of how much disparate they are from each other with respect to their conceptualizations. These models can then be screened using the proposed approach and sorted such that models with better skills will be together. Configurations of the models that obtain similar ranks can be studied to highlight model features that are common among the model sub-set. For example, in the present study the top 10 and the bottom 10 models were classified on the basis of the particular feature-states their configurations contain. All top 10 models contained feature state V12 (variable thickness of the bottom of L1); no top models contained state V11 (uniform thickness of the bottom of L1). This suggests that models with former state should be retained for further exploration that could include conventional approaches such as ensemble averaging of the model-simulated values, or further improving the numerical fit of the top-ranked model using model calibration. Hence, the proposed approach is recommended as a robust add-on filter to modelers dealing with model uncertainty to further refine their modeling solution.

The proposed approach can also be modified to suit validation of transient or dynamic state models with a time component. With transient or dynamic state models, it is important that the model's behavior matches the real world system over the selected time period. Therefore, the pattern agreement, or the consistency in model performance over the range of the transient period is equally important as the closeness of the model fit to a particular observation (Jolliff et al. 2009). In the present approach, the observational data from across the observational time were used to form the descriptor-based ECDF. This approach collated the observed data across time, although the chronological arrangement of these data was replaced by the magnitude-based monotonically increasing ordering. This arrangement is suitable for steady-state models that were simulated here. Additional procedural steps need to be taken if the objective of the exercise is changed to generate and assess the replicative validity of multiple, transient-state models with regard to pattern agreement as the primary criterion for model performance, along with the closeness of model fit.

For example, time dependent type of observed data could be time averaged to obtain a steady state value that could then be used to construct the ECDF; however, it would be inappropriate for the purposes of pattern agreement to summarize the behavior of the observed

quantity into a scalar point-value. An alternative could be to divide a continuous time period of the observations into multiple discrete time steps and generate an area metric value for each time step, assuming each step is a steady state. The resultant multiple area metric values can then be collated into an ECDF for the model (say, $ECDF_{time}$) similar to the $ECDF_{model}$ derived from the 133 area metric values generated above. In short, an area metric based multi-model validation assessment for transient state models would entail an additional area metric calculation step (Figure 5.1-b) to the process for the steady-state validation assessment above (Figure 5.1-a).

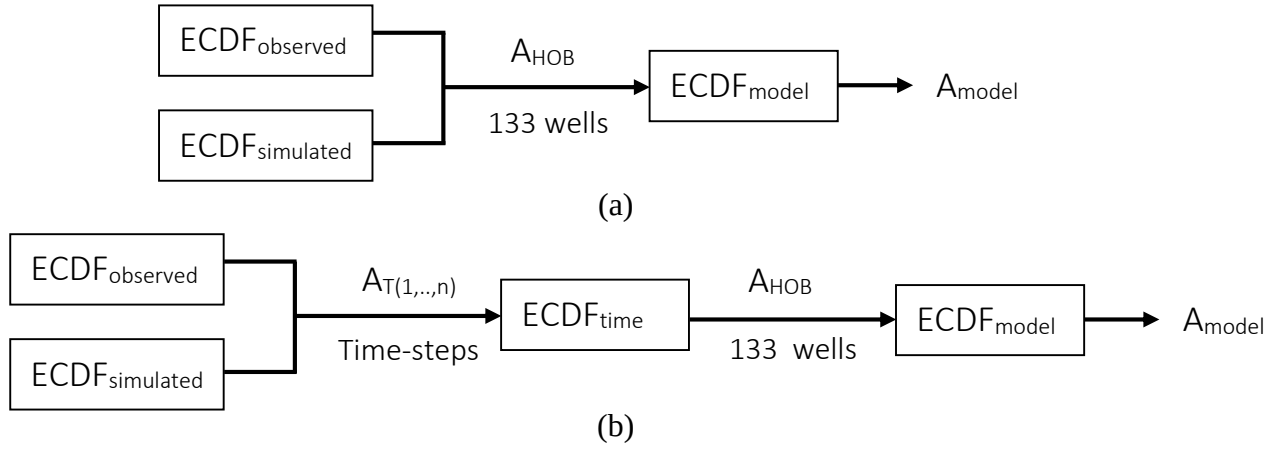


Figure 5.1: Flow chart showing process of area metric-based multi-model validation assessment for (a) steady-state models, and that for (b) dynamic models; T= number of time-steps from 1…n

The proposed approach can also be modified to suit the validation requirements of specific modeling problems. For example, the proposed approach can be expanded to include other types of observational data to assess models' validity over multi-dimensional data. The following figure conceptualizes a multi-dimensional validation assessment using the area metric approach. Here, the validity of a groundwater flow simulation model is assessed using two types of observational data – groundwater heads and streamflow measurements. The model is simulated to generate simulated head values and streamflow volumes that can then be compared with their respective real world observations to calculate the area metric scores for both types of observational data (dimensions). These scores can then be used as co-ordinates to plot the model in a two-dimensional area metric space where the origin represents the position of the ideal model ($A_{head} = A_{streamflow} = 0$) (Figure 5.2).

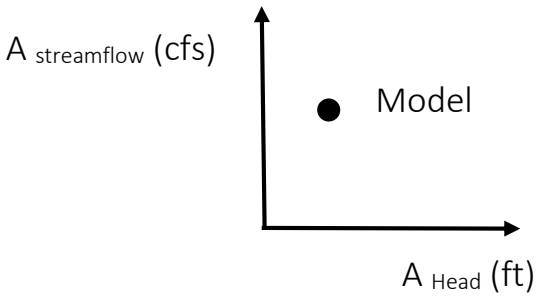


Figure 5.2: Conceptual plot for a 2-dimensional area metric-based multi-model validation

The proposed approach can also be modified to assess the validity of a contaminant transport model where the model would simulate the advective-dispersive flow of contaminants and their interaction with the surrounding material, the groundwater, and with the surface water. For example, suppose the landfill model evolves into a contaminant transport model where concentrations of three chemical species – chloride, ammonia, and iron – are measured at monitoring well locations. The model-simulated results for these 3 species can then be compared with the observational data to calculate three area metric scores of the model, one for each species. Figure 5.3 shows the positioning of the model in a three dimensional space where the origin indicates the position of the ideal model (where $A_{Cl} = A_{Fe} = A_{NH3} = 0$).

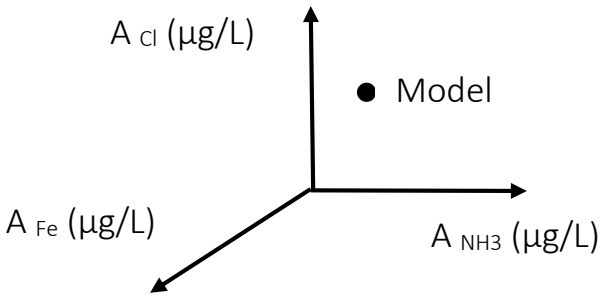


Figure 5.3: Conceptual plot for a 3-D area metric-based multi-model validation

Both of these modified approaches are data intensive because observational data are needed to accurately simulate the additional features (streams) and interaction (advection, diffusion, chemical reactions of different chemical species, sorption etc.) and to compare the simulated results with the observed data. However, these modified approaches create additional and better scrutiny to the validation process that would help reduce the type I and type II risks mentioned previously.

5.8. Summary

In summary, the area metric-based multi-model validation assessment enables explicit incorporation of model uncertainty by generating multiple model variants. Instead of adopting a binary, valid/invalid, approach, the proposed approach assessed the degree of these models' validity. The validation assessment was performed over a range of observed data to generate a composite measure of the models' performances. The proposed approach is more realistic than the conventional hypothesis testing because it acknowledges that exact correspondence between observed data and simulated output is difficult to find, given model uncertainty. In addition, this approach is flexible given that different types of observed data (multi-dimensional testing) can be incorporated in the testing scheme to increase the scrutiny of the validation assessment. This increases the rigor of the model validation process that, in turn, would increase the confidence on the validity of the chosen model(s). Also, the proposed approach can be modified according to the needs (steady-state or transient-state) and the objectives (descriptive or predictive) of the modeling exercise. This flexibility of proposed approach makes its application generalizable.

The ability to accommodate model uncertainty, flexibility, improved rigor, and generalizability makes the proposed approach utilizable in modeling exercises elsewhere. Modelers can customize the methodology of the proposed approach either by changing the number of acknowledged model uncertainties, the number and the configurations of the models that are to be tested, the type and quantities of the observational data used to develop the ECDFs, and/or the steps involved in the calculation of the area metric to suit the requirements of their respective modeling exercises. As for the case study, the proposed approach was applied to select, from a model space, better representations of the groundwater flow system in the vicinity of the Brookhaven landfill on the basis of their replicative validity over a range of groundwater heads. The solution(s) of this inverse problem can be utilized to develop modeling solution(s) for the predictive purposes.

Chapter 6

Conclusions

Models offer a powerful tools for testing theories about our understanding of the real world systems, for collating, organizing, and storing observations and empirical data, and for predicting the future behavior of the system and also the effectiveness of a design or policy change. Therefore, modelers need to build representations of the real world that have performance characteristics near to that of the real world.

Developing representative models for real world systems, such as groundwater flow systems, is challenging given these systems are sub-surface, complex, distributed, heterogeneous, and non-linear. They consist of a number of interacting elements and mechanisms that may have different impacts on the system behavior. The exact configuration of these systems is seldom known. Samples are usually taken to learn about these systems, but not all components can be directly measured. The modeler has to accommodate model uncertainty: conceptual, input, and parameter. The magnitude of model uncertainty may vary from complete determinism to total ignorance. Model uncertainty can be epistemic uncertainty that arises because of lack of knowledge about the system that is being modeled, or aleatory uncertainty that arises because of the inherent variability in the system. Model uncertainty makes it difficult to establish the representativeness of a model and limits modeling's decision support abilities.

It has been proposed that multiple models should be developed and evaluated as a remedy to address aspects of uncertainty. In this way, a modeler is free to test alternate model conceptualizations and is not restricted to validate, or defend, anyone particular model. The process of evaluation of models' representativeness is referred to as validation. A common form of validation is replicative validation, where a model's goodness-of-fit is measured by quantifying the level of agreement between the observed data and the corresponding model simulated values. Root Mean Squared Error (RMSE) is the commonly used performance measure for groundwater model validation. Different approaches to replicative validation, such as information criteria-based model selection, multi-model averaging, multi-objective optimization, and Generalized Likelihood Uncertainty Estimation (GLUE) have been used in the past for multi-model analysis. These validation methods assume that the simulated and observed data are deterministic quantities that are devoid of uncertainty. In reality, the observed data may be of uncertain quality, may be a snapshot, and may be an inadequate representation of reality. Also, exact correspondence between observed and simulated data is difficult to find because of incompatibilities in scale of input information, uncertainty or missing data, inconsistent

observations, lack of location specific information, use of surrogate data, poorly defined correlations between the input information and the observed data, or incompatible discretization levels. Important values such as the hydraulic conductivity of an aquifer are mathematically derived and cannot be directly measured. Parameterization may underrepresent the heterogeneity and distributed nature of reality. Hence, the traditional approaches of multi-model validation assessment do not achieve their ends. Therefore, better alternative is to emphasize the consistency of model behavior over the observed range of real world conditions, rather than trying to obtain the closest fit to an uncertain set of observed data.

Here, an alternative validation assessment approach, the area metric (Ferson et al. 1998) was used for multi-model validation assessment. The area metric quantifies the level of difference between observed and simulated data by calculating the area enclosed between the distributions of data. Specifically, it is the difference between the empirical cumulative probability distribution (ECDF) derived from the observed data ($ECDF_{observed}$) and the ECDF derived from the simulated outputs ($ECDF_{simulated}$). The area metric based multi-model validation assessment was used to assess the replicative validity of multiple model variants of a groundwater flow simulation model developed for the municipal landfill site in the Town of Brookhaven in Suffolk County, New York. It was hypothesized that the area metric based multi-model validation assessment facilitates a robust multi-model analysis that selects those model variants that are better representations of the real world groundwater flow system, with respect to their replicative validity.

Divergent opinions exist with regard to the geologic framework of the study area. Intermittent physical and qualitative data, spanning over three decades, about the groundwater and surface-water are available. Field-measurements and historical information for the aquifer characteristics such as the hydraulic conductivity show considerable heterogeneity. Hence, 288 model variants of a based model were generated to represent uncertainty in select model features. Each model was a unique combination of fixed and variable features. The uncertainty in variable features was represented by either two or three states of the variable feature. Seven variable features represented epistemic uncertainty. One variable feature represented the aleatory uncertainty. The models were simulated using Visual MODFLOW v. 4.2. A total of 133 head observation wells were included in the analysis.

The $ECDF_{observed}$ for each well was derived from three data points: the minimum, median, and maximum head observation values. The $ECDF_{simulated}$ was derived from three simulation data points derived by simulating each model variant for three groundwater conditions: low, median, and high groundwater levels. The value of the area metric for the well (A) was calculated from $ECDF_{observed}$ and its corresponding $ECDF_{simulated}$ of a given well; this process was repeated for each of the 133 wells and for all models. In the second step, the area metric values generated for each well were collated to develop a model ECDF ($ECDF_{model}$) for each model. Each $ECDF_{model}$ was compared to the ECDF of a hypothetical reference model assumed to have a perfect overlap between the observed and the simulated data for each well (A = 0 for all wells); this generated the value of the model area metric (A*) for each model. Models with low A* value were deemed to have higher replicative validity than the models with high A* value. Twenty-three models could not be included in the analysis due to their simulations' abnormal termination. A simple subtraction of the adjacent simulated values (high - median, median - low) indicated that the logical ordering of the simulated values was violated in 65 models. These models were excluded, leaving 200 model variants for further analyses.

These 200 models were ranked according to ascending order of magnitude of their model area metric (A*) scores. The minimum A* value (1.257 feet) was obtained for model #178, while the maximum A* value (2.926 feet) was obtained for model #169. Models with smaller A* values were considered as better representations of the groundwater flow system in the vicinity of the Brookhaven landfill. The ECDFs of these 200 models showed horizontal as well as vertical dispersion that, in turn, indicated that the agreement between the model simulated values and the head observations differed across the 200 models as well as across the 133 wells. The well area metric (A) values were alternatively depicted in the form of boxplots superimposed over by the model area metric (A*) values for each model. This figure highlighted the distinctly low A* values for the top 5-10 models, the increasing interquartile ranges from the top to the bottom models, the strong correspondence between the A* values and the boxplot medians, and the presence of singular outliers.

The means for the well area metric (A) values calculated from A values for the 133 wells of all 200 models, as well as for the top 10 models and for the bottom 10 modes. The means of A values for the top 10 models were generally smaller compared to the means of A values calculated for all models and these, in turn, were smaller compared to the means of A values

calculated for the bottom 10 models. The differences among these groups were found to be statistically significant using ANOVA and pairwise t-tests. The means of well area metric (A) values for all 200 models varied from 0.586 feet for well S72160 to 4.438 feet for well S96202. Larger A values were found near the southern edge of the landfill to the upper reaches of the downgradient streams. The mean A values increased with the increase in the measuring point elevations.

The configurations of the top 10 and the bottom 10 models were qualitatively analyzed. This analysis indicated that the model configurations of the top and the bottom models were distinctive with varying prevalence of variable features and states in these configurations. This analysis was extended for the configurations of all 200 models. The boxplot and the unbalanced one factor ANOVA analyses indicated that feature-states V12 (uniform thickness of layer 1), V21 (position of the bottom of layer 2 close to the bottom of layer 1), V32 (3-zone configuration of the potentially semi-confining unit), and V61 (high preamble configuration of the Upper Glacial aquifer's conductivity) were found to be statistically more frequent in those models with lower A* values. It was concluded that local heterogeneities in the models' hydrogeologic settings and aquifer characteristics affected the geospatial distribution in the A values. Also, it was concluded that the significance or not of these states is dependent on the choice of the system response quantity.

The association between the A* values and the corresponding RMSE values was noticeable ($R^2 = 0.657$). Two sensitivity analysis experiments were conducted. It was found that the A* values were generally lower and there was a noticeable change in the model rankings when the range of the observed data descriptors was changed from the original "min-max" range to the quartile range. A strong linear association ($R^2 = 0.687$) was found between the quartile-based and the "min-max"-based A* values. Increasing the ECDF resolution from 3 to 5 data points increased the A values for most wells for most models.

The area metric based multi-model validation assessment approach ("the proposed approach") is better suited for a multi-model validation assessment compared to other approaches adopted in the model validation literature on the basis of the following salient features.

First, model uncertainty was explicitly represented using multiple model variants of a base landfill model rather than using a singular model. The modeler's subjective understanding

of the groundwater system may be incomplete or flawed if the observational data that describe the system are inadequate or uncertain. In such situation, more than one conceptualization can be developed and tested. Therefore, developing and testing multiple models is a better alternative given model uncertainty than treating a singular model as error-free. Here, 288 model variants of a base landfill model were generated and evaluated. Each model represented a unique combination of different states of uncertain model features combined with fixed model features.

Second, the proposed approach offers a more pragmatic alternative to the traditional schemes based on hypothesis testing where a model can either be validated (verificationism) or not validated but can be invalidated (falsificationism). This binary acceptance or rejection of a model's validity is not achievable given the uncertainty in our understanding of the real world system. Instead, the proposed approach assessed the degree of models' validity, that is, the level of agreement between observed and simulated values. The area metric values did not ratify or refute the validity of any particular model.

Third, in the proposed approach it is acknowledged that uncertainty reduction or elimination is difficult to achieve given the potential for confounding of model errors and the difficulty in apportioning uncertainty to its sources in a complex and heterogeneous model. The proposed approach offers a more pragmatic alternative by classifying uncertainties into reducible (epistemic) and irreducible (aleatory) classes. This segregation allows assessment of how each uncertainty type impacts models' performances that, in turn, could streamline the data gathering efforts in mode gainful directions. Here, it was found that the uncertainty in models' hydrogeologic settings and aquifer characteristics affected their performances. Future data gathering efforts can be focused on bringing more clarity over these aspects.

Fourth, a blind assessment was conducted of the degree of multiple models' representation using the proposed approach instead of calibrating or updating a singular model to obtain a better numerical fit with the observed data. A blind assessment reduces the chances of model over-fitting, the model configuration is not tuned or updated to obtain an exact fit with the calibration data set. Here, each of the 200 models retained their initial model make-up throughout. These models were ranked on the basis of their model area metric (A*) values, without subjecting them to tuning/updating.

Fifth, the proposed approach acknowledges that exact correspondence between observed data and simulated output is difficult to find. Matching model results with a singular, snapshot

representation of the observed data is generally thought to limit the model's applicability to other data conditions given the epistemic and aleatory uncertainty associated with these data. Hence, the validity of multiple models was assessed over a range of observed groundwater head data represented as their empirical cumulative distribution functions.

Sixth, the area metric used in the proposed approach offers a flexible approach in assessing the model performance. The observed and the simulated data can be collapsed into a cumulative distribution function allowing model assessment over a range of these data. Here, three states of the groundwater conditions (high, median, and low), and their corresponding simulated results, were collated into an ECDF for each of the 133 wells to calculate the well area metric (A). Subsequently, these 133 A values were then collated into a model-ECDF to calculate the model area metric (A*) for each of the 200 models against a reference model's ECDF. Also, the area metric is non-parametric and independent of the number of the simulation and experimental samples. The area metric is also mathematically well-behaved and is expressed in the same units as the observed data.

Seventh, the flexibility of the proposed approach can be generalized to other modeling exercises that may have similar or different modeling objectives or systems under study. The dimensionality of the system response quantity can be increased to include other observational data, such as streamflow volumes or water quality indicators in addition to the groundwater heads, to assess the level of agreement between these data and the model-simulated values. Additional procedural step can be included in the proposed approach to accommodate pattern matching that is important in cases of transient state models. Here, the solution(s) obtained for the inverse, groundwater flow simulation, problem can be used to find solution for the forward, contaminant fate and transport, problem.

The validation assessment using the proposed approach assessed the replicative validity of the model. Additional assessment will be needed to expand the scope of validation exercise to other areas of the model such as conceptual validity or predictive validity. Also, the validation assessment using the proposed approach is relative given the model space may not be exhaustive. It is important to remember that the utility of the proposed approach is best realized with realistic understanding of its application domain.

The multi-model validation assessment using the area metric approach is firmly rooted in the pragmatic realism about how models are built and tested. Instead of treating the model

conceptualization as immutable, this approach incorporates epistemic and aleatory model uncertainties into the validation assessment in the form of multiple model variants. This approach assesses the degree of models' validity, instead of resorting to binary model validation, or invalidation. This assessment is made over a range of observed data to avoid model overfitting to a particular system state. This approach is flexible, scientifically rigorous, and is generalizable for other modeling exercises to suit their requirements. 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 validation 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.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in Statistics*, Vol. 1, Foundations and Basic Theory, S. Kotz and N. Johnson (Ed). Springer-Verlag, New York, NY: 610-624.

Anderson, D.R. and K.P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management* 66, no. 3: 912-918.

Anderson, M. and W. Woessner. 1992. *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*, Vol. 4, Academic Press, New York, NY: 2.

Anderson, M. and W. Woessner. 1992. ibid.: 3.

Anderson, M. and W. Woessner. 1992. ibid.: 4-5.

Anderson, M. and W. Woessner. 1992. ibid.: 15.

Anderson, M. and W. Woessner. 1992. ibid.: 20.

Anderson, M. and W. Woessner. 1992. ibid.: 45.

Anderson, M. and W. Woessner. 1992. ibid.: 69.

Anderson, M. and W. Woessner. 1992. ibid.: 226.

Aphale, O. and D.J. Tonjes. 2010. Recharge and head: preliminary findings using a long-term data set. Expanded abstract, presented at the Seventeenth Conference on Geology on Long Island and Metropolitan New York. Available at http://www.geo.sunysb.edu/lig/Conferences/abstracts-10/10-program.htm. Accessed on March 22, 2012.

Aphale, O. and D.J. Tonjes. 2013. Geology in the vicinity of the Town of Brookhaven (Suffolk County, New York) landfill. Waste Reduction and Management Institute, School of Marine and Atmospheric Sciences, Stony Brook University: 100 p.

Aquavevo, LLC .2010. Groundwater Modeling System (GMS). Available at http://www.aquaveo.com/software/gms-groundwater-modeling-system-introduction.

Aronson, D.A., J.B. Lindner, and B.G. Katz B.G. 1983. Geohydrology of the Meadowbrook artificial recharge site at East Meadow, Nassau County, New York. *USGS Water Resources Investigations Report* 82-4084: 44 p.

ASTM D5490-93 .2002. Standard guide for comparing ground-water flow model simulations to site-specific information. ASTM International, West Conshohocken, PA.

Balci, O. 1998. Verification, validation, and accreditation. *Proceedings of the 1998 Winter Simulation Conference*, D.J. Mederios, E.F. Watson, J.S. Carson, and M.S, Manivannan, (Ed), IEEE, Piscataway, NJ: 41-48.

Barlas, Y. 1996. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review* 12, no. 3: 183-210.

Barlas, Y. and S. Carpenter. 1990. Philosophical roots of model validation: two paradigms. *System Dynamics Review* 6, no.2: 148-166.

Beven, K. J. 2001. Dalton medal lecture: how far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences* 5, no.1: 1-12.

Beven, K. 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* 320, no. 1: 18-36.

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

Beven, K. and A. Binley. 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes* 6, no. 3: 279-298.

Beven, K. and J. Freer. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* 249, no. 1: 11-29.

Binley, A., K. Beven, and J. Elgy. 1989. A physically based model of heterogeneous hillslopes: 2, effective hydraulic conductivities. *Water Resources Research* 25, no. 6: 1227-1233.

Brassington, R. 2013. Field hydrogeology. *Geological Society of London Handbook Series*, London, UK: 80-81.

Brassington, R. 2013. ibid.: 75.

Bredehoeft, J. 2003. From models to performance assessment: the conceptual problem. *Ground Water* 41, no. 5: 571-577.

Bredehoeft, J. 2005. The conceptualization model problem-surprise. *Journal of Hydrogeology* 13, no. 1: 37-46.

Bredehoeft, J. 2012. Modeling groundwater glow – the beginnings. *Ground Water* 50, no. 3: 324-329.

Brugnach, M., A. Tagg, F. Keil, and W.J. de Lange. 2007. Uncertainty matters: computer models at the science – policy interface. *Water Resources Management* 21, no. 7: 1075-1090.

Busciolano, R. 2002. Water table and potentiometric surface altitudes of the Upper Glacial, Magothy, and Lloyd aquifers on Long Island, New York, in March-April 2000, with a summary of hydrogeologic conditions. *USGS Water Resources Investigations Report* 09-4165, 2nd Edition: 17 p.

Busciolano, R., J. Monti (Jr.), A. Chu. 1998. Water table and potentiometric-surface altitudes of the Upper Glacial, Magothy, and Lloyd Aquifers on Long Island, New York, in March-April 1997, with a summary of hydrogeologic conditions. *USGS Water Resources Investigations Report* 98-4019: 17 p.

Buxton, H.T. and D. Smolenksy. 1999. Simulation of the effects of development of the groundwater flow system of Long Island, New York. *USGS Water Resources Investigations Report* 98-4069: 57 p.

Buxton, H.T. and E. Modica. 1992. Patterns and rates of ground water flow on Long Island, NY. *Ground Water* 30, no. 6: 857-866.

Buxton, H.T. and T. Reilly. 1985. Effects of sanitary sewering on groundwater levels and streams in Nassau County, NY, part 1, geohydrology, modeling strategy, and regional evaluation. *USGS Water Resource Investigations Report* 82-4045: 45 p.

Carle, S.F. 1999. T-PROGS: Transition Probability Geostatistical Software v. 2.1. Hydrologic sciences graduate group: University of California, Davis: 84 p.

Chamberlin, T. C. 1965. The method of multiple working hypotheses. *Science* 148, no. 3671: 754-759.

Cashin Associates. 2002. Carmans River environmental assessment, Suffolk County, NY. Cashin Associates, Hauppauge, NY: Paged in sections.

Dan Johnson, Personal Communication, April 6, 2012.

Darcy, H. 1856. *Les fontaines publiques de la ville de Dijon*, Paris, Victor Dalmont.

DeLaguna, W. 1963. Geology of Brookhaven National Laboratory and vicinity, Suffolk County, NY. *USGS Bulletin* 1156-A: 35p.

Demarty, J., C. Ottle, I. Braud, A. Olioso, J. Frangi, L. Bastidas, and H. Gupta. 2004. Using a multi-objective approach to retrieve information on surface properties used in a SVAT model. *Journal of Hydrology* 287, no. 1: 214-236.

Doherty, J. 2011. Modeling: picture perfect or abstract art? *Ground Water* 49, no. 4: 445-455.

Doherty, J. and R.J. Hunt. 2010. Approaches to highly parameterized inversion - a guide to using PEST for groundwater-model calibration. *USGS Scientific Investigations Report* 2010–5169: 59p.

Doriski, T.P. 1987. Potentiometric-surface of the water-table, Maogthy, and Lloyd aquifers on Long Island, New York, in 1984. *USGS Water-Resources Investigations Report* 86-4189: 8 sheets.

Doriski, T.P. and F. Wilde-Katz. 1983. Geology of the "20-Foot Clay" and Gardiners Clay in southern Nassau and southwestern Suffolk Counties, Long Island, New York. *USGS Water Resources Investigations* 82-4056: 21 p.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society*, Series B (Methodological): 45-97.

Duan Q., S. Sorooshian, and V. Gupta. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* 28, no. 4: 1015-1031.

Dvirka and Bartilucci Consulting Engineers (D&B). 2010. Leachate plume characterization report for the Town of Brookhaven landfill, Suffolk County, New York. Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 1994a. Part 360 Hydrogeological investigation report, Brookhaven landfill expansion – Cell 5, Part IV (of landfill application). Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 1994b. Part 360 Hydrogeological investigation report, Brookhaven landfill expansion – Cell 5, Part IV, Appendices A through C. Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 1994c. Part 360 Hydrogeological investigation report, Brookhaven landfill expansion – Cell 5, Part IV, Appendices D through N. Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 1996. Part 360 Closure investigation report, Brookhaven landfill Cells 1-4, Volume I. Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 2001. Part 360 Hydrogeological investigation report, Brookhaven landfill expansion – Cell 6. Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 2010. Data summary report and leachate plume monitoring plan, Town of Brookhaven landfill, Suffolk County, New York. Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 2011. Leachate plume characterization report for the Town of Brookhaven landfill, Suffolk County, New York. Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY: Paged in sections.

Dvirka and Bartilucci. 2012. Stream flow (discharge) measurements, Dvirka and Bartilucci Consulting Engineers (D&B), Woodbury, NY. Personal communication, April 10, 2012.

Eckhardt, A.V. and E.J. Wexler. 1986. Groundwater movement in the Upper Glacial aquifer in the Manorville area, Town of Brookhaven, Long Island, New York, in November 1983. *USGS Water Resources Investigations Report* 85-4035: 12 p.

Efstratiadis, A. and D. Koutsoyiannis. 2010. One decade of multi-objective calibration approaches in hydrological modeling: a review. *Hydrological Sciences Journal* 55, no. 1: 58-78.

Engelhardt, I., J.G. De Aguinaga, H. Mikat, C. Schuth, O. Lenz, and R., Liedl. 2012. Complexity versus simplicity: an example of groundwater model ranking with Akaike Information Criterion. *Hydrology and Earth System Sciences Discussions* 9, no. 8: 9687-9714.

Environmental Solutions Inc. 2015. Groundwater Vistas. Available at http://www.groundwatermodels.com/ESI_Software.php.

Fanning, Phillips, and Molnar. 1986. Geohydrological investigation of the regional resource recovery ashfill site at Yaphank. Fanning, Phillips, and Molnar Engineers, Plainview, New York: 14 p. + Appendices + plates.

Ferson, S. 2001. Probability bounds analysis solves the problem of incomplete specification in probabilistic risk and safety assessments. *Risk-Based Decision making in Water Resources* IX: 173-188.

Ferson, S., W.L. Oberkampf, and L. Ginzburg. 2008. Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering* 197, no. 29: 2408-2430.

Ferson, Scott, and W. T. Tucker. 2003. Reliability of Risk Analyses for Contaminated Groundwater. *Groundwater quality modeling and management under uncertainty*. ASCE, 2003.

Ferson, S. and W.T. Tucker. 2006. Sensitivity analysis using probability bounding. *Reliability Engineering & System Safety*, The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) 91, no. 10-11: 1435–1442.

Fetter, C. 2001. Applied Hydrogeology, 4th Edition, Prentice Hall, New Jersey: 3.

Fetter, C. 2001. ibid.: 79.

Fetter, C. 2001. ibid.: 96.

Fetter, C. 2001. ibid.: 101.

Fetter, C. 2001. ibid.: 224.

Fetter, C. 2001. ibid.: 515.

Fetter, C. 2001. ibid.: 516.

Forrester, J. 1961. Judging Model Validity, in *Industrial Dynamics*, Pegasus Communications, Waltham, MA: 115-130.

Forrester, J. and P.M. Senge. 1980. Tests for Building Confidence in System Dynamics Models, in *System Dynamics*, A.A. Legasto (Jr.), J., Forrester, J.M., Lyneis (Ed), TIMS Studies in the Management Sciences, North-Holland, New York, NY, Vol. 14: 201-228.

Franke, O.L. and P. Cohen. 1972. Regional rates of groundwater movement on Long Island, New York. *USGS Professional Paper 600-B*: B205 - B209.

Franke, O.L. and N.E. McClymonds. 1972. Summary of the hydrologic situation on Long Island, New York, as a guide to water management alternatives. *USGS Professional Paper 627-F*: 59.

Freer, J.E., K.J. Beven, and B., Ambroise. 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. *Water Resources Research* 32, no. 7: 2161-2173.

Garber, M.S. 1986. Geohydrology of the Lloyd Aquifer, New York. *USGS Water Resources Investigations Report* 85-4159: 36 p.

Gass, S.I. 1983. Decision-making models: validation, assessment, and related issues for policy analysis. *Operations Research* 31, no. 4: 603-631.

Gerathy and Miller. 1985. Results of hydrogeologic site studies at the Yaphank and Moriches sites. Gerathy and Miller, Syosset, New York: 16 p.

Gupta, H., S. Sorooshian, and P. Yapo. 1998. Towards improved calibration of hydrologic models: multiple and incommensurable measures of information. *Water Resources Research* 34, no. 4: 751-763.

Gureghian, A.B., D.S. Ward and R.W. Cleary. 1980. A finite-element model for the migration of leachate from a sanitary landfill in Long Island, New York, part I, theory. *Water Resources Bulletin* 16, no. 5: 900-906.

Harbaugh, A. and R. Getzen. 1977. Stream simulation in an analog model of the groundwater system on Long Island, New York. *USGS Water Resources Investigations* Report 77-58: 15 p.

Harbaugh, A.W. 1990. A computer program for calculating subregional water budgets using results from the USGS modular three-dimensional finite-difference ground water flow model. *USGS Open File Report* 90-392: 49 p.

Harbaugh, A.W. 2005. MODFLOW-2005, The USGS modular ground-water model – the ground-water flow process: *USGS Techniques and Methods,* Book 6, Chapter A16: Variously Paginated.

Harbaugh, A.W., E.R. Banta, M.C. Hill, and M.G. McDonald. 2000. MODFLOW-2000, the USGS modular ground-water model: user guide to modularization concepts and the ground-water flow process. *USGS Open-File Report* 00-02: 121 p.

Harbaugh, A.W. and M.G. McDonald. 1996. User's documentation for MODFLOW–96, an update to the USGS modular finite-difference ground-water flow model. *USGS Open-File Report* 96–485: 56 p.

Harte, P. 1994. Comparison of vertical discretization techniques in finite-difference models of ground-water flow; example from a hypothetical New England setting; *USGS Open-File Report* 94-343: 24 p.

Hill, M.C. 1990. Preconditioned Conjugate-Gradient 2 (PCG2), A computer program for solving ground-water flow equations. *USGS Water-Resources Investigations Report* 90-4048: 43 p.

Hill, M.C., E.R. Banta, A.W. Harbaugh, and E.R. Anderman. 2000. MODFLOW-2000, the USGS modular ground-water model - user guide to the observation, sensitivity, and parameter-estimation processes and three post-processing programs. *USGS Open-File Report* 00-184: 210 p.

Hills, R. G. 2006. Model validation: model parameter and measurement uncertainty, *Journal of Heat Transfer* 128, no.4: 339-351.

Hoeting, J. A., D. Madigan, A.E. Raftery, C.T. Volinsky. 1999. Bayesian model averaging: a tutorial. *Statistical Science*: 382-401.

Isbister, J. 1962. Relation of fresh-water to salt water at Center Island, Nassau County, New York. *USGS Professional Paper* 450-E, article 226: E154-E156.

Jansen, J. 2003. Parameter and uncertainty estimation in groundwater modeling. PhD Thesis, Aalborg University, Aalborg, Denmark: 143 p.

Janssen, P.H.M. and P.S.C. Heuberger. 1995. Calibration of process-oriented models. *Ecological Modelling* 83, no. 1: 55-66.

Johnson, D., Personal communication, April 6, 2012.

Jolliff, J. K., J. C. Kindle, I. Shulman, B. Penta, M. A. Friedrichs, R. Helber, and R. A. Arnone. 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems* 76, no. 1: 64-82.

Karnaki, D., H. Kushwaha, A. Verma, and S. Ajit. 2009. Uncertainty analysis based on probability bounds (P-box) approach in probabilistic safety assessment. *Risk analysis*, 29, no. 5: 662-675.

Kashyap, R.L. 1982. Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4, no. 2: 99-104.

Kimmel, G.E. and O.C. Braids. 1980. Leachate plumes in groundwater from Babylon and Islip landfills, Long Island, New York. *USGS Professional Paper* 1085: 38 p.

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.

Konikow, L. F. and J.D. Bredehoeft. 1992. Ground-water models cannot be validated. *Advances in Water Resources* 15, no. 1: 75-83.

Koppelman, Lee E. 1978. The Long Island comprehensive waste treatment management plan, Volumes I and II. Long Island Regional Planning Board (LIRPB), Hauppauge, NY.

Koszalka, E.J. 1984. Geohydrology of the northern part of the Town of Brookhaven, Suffolk County, New York. *USGS Water Resources Investigations Report* 83-4042: 37 p.

Krulikas, R.K. 1986. Hydrologic appraisal of the Pine Barrens, Suffolk County, New York. *USGS Water Resources Investigations Report* 84-4271: 53 p.

Lahsen, M. 2005. Seductive simulations? Uncertainty distribution around climate models. *Social Studies of Science* 35, no. 6: 895-922.

Lane, S.N. and K.S. Richards. 2001. The "validation" of hydrodynamic models: some critical perspectives, in *Model Validation: Perspectives in Hydrological Science*, M.G. Anderson and P.D. Bates (Ed), Wiley and Sons, New York, NY: 413-438.

Law, A. and W.D. Kelton. 2000. Validation of simulation models, simulation modeling and analysis, McGraw Hill, 3rd Edition, Boston, MA: 264.

Law, A.M. 2005. How to build valid and credible simulation models, *Proceedings of the 2005 Winter Simulation Conference*, M.E. Kuhl, N.M. Steiger, F.B. Armstrong, and J.A. Joines (Ed). IEEE, Piscataway, NJ: 24-32.

Legates, D.R. and G.J. McCabe (Jr.). 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35, no. 1: 233-241.

Lindner, J.B. and T.E. Reilly. 1983. Analysis of three tests of the unconsolidated aquifer in southern Nassau County, Long Island, New York. *USGS Water Resources Investigations* Report 82-4021: 46 p.

Loague, K. and R.E. Green. 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. *Journal of Contaminant Hydrology* 7, no. 1: 51-73.

Lockwood, Kessler and Bartlett, Inc. 1994. Groundwater plume remediation: aquifer test for evaluating groundwater capture and recovery systems at the Brookhaven Landfill. Lockwood, Kessler and Bartlett Consulting Engineers, Syosset, New York.

Mantovan, P. and E. Todini. 2006. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. *Journal of Hydrology* 330, no. 1: 368-381.

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.

McDonald, M. G. and A.W. Harbaugh. 1988. A modular three-dimensional finite-difference ground-water flow model. *USGS Techniques of Water-Resources Investigations*, Book 6, Chapter A1, 588 p.

McDonald, M.G. and A.W. Harbaugh. 1984. A modular three-dimensional finite-difference ground-water flow model. *USGS Open-File Report* 88-482: A1.

Mercer, J. W. and C.R. Faust. 1981. Ground-water modeling. National Water Well Association, Worthington, Ohio.

Miller, J. and R. Fredrick. 1969. The precipitation regime of Long Island, New York. *USGS Professional Paper* 627-A: 21 p.

Monti, J. (Jr) and R. Busciolano. 2009. Water-table and potentiometric-surface altitudes in the Upper Glacial, Magothy, and Lloyd aquifers beneath Long Island, New York, March-April 2006. *USGS Scientific Investigations* Map 3066.

Monti, J. (Jr), M. Como, and R. Busciolano. 2013. Water-table and potentiometric surface altitudes in the Upper Glacial, Magothy, and Lloyd aquifers beneath Long Island, New York, April-May 2010. *USGS Scientific Investigations Map* 3270: 4 sheets.

Morel-Seytoux, H.J. 2001. Groundwater, in *Model Validation: Perspectives in Hydrological Science*, M.G. Anderson and P.D. Bates (Ed), Wiley and Sons, New York, NY, 293-323.

Morgan, M.G. and M. Henrion. 2006. Uncertainty: A guide to dealing with uncertainty. In *Quantitative Risk and Policy Analysis*, Cambridge University Press, New York: 50.

Morgan, M.G. and M. Henrion. 2006. ibid.: 56.

Morgan, M.G. and M. Henrion. 2006. ibid.: 60.

Morgan, M.G. and M. Henrion. 2006. ibid.: 63.

Morgan, M.G. and M. Henrion. 2006. ibid.: 74.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, and T.L. Veith. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *American Society of Agricultural and Biological Engineers* ISSN 0001-2351: 885-900.

Neuman, S. 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment* 17, no. 5: 291-305.

Neuman, S. and 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.

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.

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* 75, no. 3: 333-357.

Oberkampf, W. L., T.G. Trucano, and C. Hirsch. 2004. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews* 57, no. 5: 345-384.

Oberkampf, W. L. and T.G. Trucano. 2002. Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences* 38, no. 3: 209-272.

Oberkmapf, W. and C. Roy. 2010. Verification and validation in scientific computing, Cambridge University Press, New York: 98.

Oberkmapf, W. and C. Roy. 2010. ibid.: 374.

Oberkmapf, W. and C. Roy. 2010. ibid.: 401.

Oreskes, N. 1998. Evaluation (not validation) of quantitative models. *Environmental Health Perspectives* 106, no. 6: 1453-1460.

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.

Palmer, A., M. Palmer, and I. Sasowsky, I. (Ed). 1999. *Karst modeling* (Vol. 5). Karst Waters Institute, Charles Town, West Virginia: 272 p.

Pearsall, K.A. and M.J. Aufderheide. 1995. Groundwater quality and geochemical processes at a municipal landfill, Town of Brookhaven, Long Island, New York. *USGS Water Resources Investigations Report* 91-4154: 45 p.

Perlmutter, N.M. and J.J. Gerathy. 1963. Geology and groundwater conditions in southern Nassau and southeastern Queens Counties, Long Island, New York. *USGS Water Supply Paper* 1613-A: 205 p.

Peterson, D.S. 1987. Groundwater recharge in Nassau and Suffolk Counties, New York. *USGS Water Resources Investigations Report* 86-4181: 19 p.

Pinder, G.F. 1973. A Galerkin finite-element simulation of groundwater contamination on Long Island, New York. *Water Resources Research* 9, no. 6: 1657-1669.

Pluhowski, E.J. and I.H. Kantrowitz. 1964. Hydrology of the Babylon-Islip Area, Suffolk County, Long Island, New York. *USGS Water Supply Paper* 1768: 119 p.

Poeter, E. P., M. C. Hill, E. R. Banta, S. Mehl, and S. Christensen. 2005. UCODE-2005 and six other computer codes for universal sensitivity analysis, calibration, and uncertainty evaluation. *USGS Technical Methods*, 6-A11: 282 p.

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

Pollock, D.W. 1994. User's guide for MODPATH/MODPATH-PLOT, Version 3: A particle tracking post-processing package for MODFLOW, the USGS finite-difference ground-water flow model. *USGS Open-File Report* 94-464: 249 p.

Popper, K. 1959. The logic of scientific discovery. Basic Books, New York, NY: 27-34.

Prince, K. 1980. Preliminary investigation of a shallow ground water flow system associated with Connetquot Brook, Long Island, New York. *USGS Water Resources Investigations* 80-47: 23 p.

Prince, K.R., O.L. Franke, T.E. Reilly. 1988. Quantitative assessment of the shallow groundwater flow system associated with Connetquot Brook, Long Island, New York. *USGS Paper* 2309: 28 p.

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

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* 29, no. 11: 1586-1597.

Refsgaard, J.C. and J. Knudsen. 1996. Operational validation and intercomparison of different types of hydrological models. *Water Resources Research* 32, no. 7: 2189-2202.

Reilly, T.E. and A.W. Harbaugh. 2004. Guidelines for evaluating ground-water flow models. *USGS Investigations Report* 2004-5038: 30 p.

Reilly, T.E., H.T. Buxton, O.L. Franke, and R.L. Wait. 1983. Effects of sanitary sewers ion ground water levels and streams in Nassau and Suffolk Counties, New York, Part 1: geohydrology, modeling strategy, and regional evaluation. *USGS Water Resources Investigations Report* 82-4045: 45 p.

Rojas, R., L. Feyen, and A. Dassargues. 2008. Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research* 44, no. 12: W12418.

Romanowicz, R. and R. MacDonald. 2005. Modeling uncertainty and variability in environmental systems. *Acta Geophysica Polonica* 53, no. 4: 401-417.

Romanowicz, R.J. and K.J. Beven. 2006. Comments on generalized likelihood uncertainty estimation. *Reliability Engineering & System Safety* 91, no. 10: 1315-1321.

Romero, V. J. 2007. Validated model? Not so fast. The need for model "conditioning" as an essential addendum to model validation. In Proceedings of the 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference.

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

Rozell, D. and S. Reaven. 2011. Water pollution risk associated with natural gas extraction from the Marcellus Shale. *Risk Analysis* 32, no. 8: 1382-1393.

Rykiel, E.J. (Jr). 1996. Testing ecological models: the meaning of validation. *Ecological Modeling* 90, no. 3: 229-244.

Sargent, R.G. 2009. Verification and validation of simulation models. Proceedings of the 2009 Winter Simulation Conference, M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin, R.G. Ingalls (Ed), *IEEE*, Piscataway, NJ: 162-175.

Schellenberger, R. 1974. Criteria for assessing model validity for managerial purposes. *Decision Sciences* 5, no. 4: 644-653.

Scorca, M.P., T.E. Reilly, and O.L. Franke. 1995. Selected hydrogeologic and water quality data from Jones Beach Island, Long Island, New York. *USGS Water Resources Investigations Report* 92-4171: 21 p.

Seaburn, G.E. and D.A. Aronson. 1974. Influence of recharge basins on the hydrology of Nassau and Suffolk Counties, Long Island, New York. *USGS Water Supply Paper* 2031: 61 p.

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

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.

Sirkin, L.A. 1986. Palynography and stratigraphy of Cretaceous and Pleistocene sediments on Long Island, New York - A basis for correlation with New Jersey coastal plain sediments. *USGS Bulletin* 1559: 44p.

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.

Smolensky, D.A., and S.M. Feldman. 1992. Geohydrology of the Bethpage-Hicksville-Levittown Area, Long Island, New York. *USGS Water Resources Investigations Report* 88-4135: 25 p.

Soren, J. 1971. Results of subsurface exploration in the mid-island area of western Suffolk County, Long Island, New York. Suffolk County Water Authority, Long Island *Water Resources Bulletin* 1: 60 p.

Spinello, A.G. and D.L. Simmons. 1992. Base flow of 10 south shore streams, Long Island, New York, 1976-85, and the effects of urbanization of base flow and flow duration. *USGS Water Resources Investigations Report* 90-4205: 34 p.

Stedinger, J. R., R.M. Vogel, S.U. Lee, and R. Batchelder. 2008. Appraisal of the Generalized Likelihood Uncertainty Estimation (GLUE) method. *Water Resources Research*, 44(12), W00B06, doi:10.1029/2008WR006822.

Steenhuis, T.S., C.D. Jackson, S.K. Kung, and W. Brutsaert. 1985. Measurement of groundwater recharge on eastern Long Island, New York, USA. *Journal of Hydrology* 79, no. 1-2: 145-169.

Sterman, J.D. 2006. Learning from evidence in a complex world. *American Journal of Public Health* 96, no. 3: 505-514.

Stone, B.D., H.W. Borns (Jr). 1986. Pleistocene glacial and interglacial stratigraphy of New England, Long Island, and adjacent Georges Bank and Gulf of Maine. *Quaternary Geochronology* 5: 39 p.

Todini, E and P. Mantovan. 2007. Comment on “on undermining the science?” by Keith Beven. *Hydrological Processes* 21, no. 12: 1633-1638.

Tonjes, D.J., Personal Communication, July 6, 2015.

Tonjes, D.J. and R.J. Wetjen. 2002. Groundwater head measurements in south-central Suffolk County, New York, 1975-2001. Expanded Abstract, presented at the Sixth Conference on Geology on Long Island and Metropolitan New York. Available at http://www.geo.sunysb.edu/lig/Conferences/abstracts_02/4_02_prog.htm.

Trucano, T. G., L.P. Swiler, T. Igusa, W.L. Oberkampf, and M. Pilch. 2006. Calibration, validation, and sensitivity analysis: what's what. *Reliability Engineering & System Safety* 91, no. 10: 1331-1357.

Tsang C.F. 1991. The modeling process and model validation. *Ground Water* 29, no. 6: 825-831.

Voorhis, C.J. 1986. Discussion of hydrogeologic zone boundaries in the vicinity of South Yaphank, Long Island, New York. Department of Planning Environment and Development, Town of Brookhaven: 16 p. + Appendices.

Voss, C. 2011. Editor's message: groundwater modeling fantasies: Part 2, down to earth. *Hydrogeology Journal* 19, no. 7: 1455-1458.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian. 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research* 39, no. 8: 1214. doi:10.1029/2002WR001746.

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

Walker, W.E., P. Harremoes, J. Rotmans, J.P.V.D. Sluijs, M.B.A.V. Asselt, P. Janssen, and M.P.K.V. Krauss. 2003. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment* 4, no. 1: 2.

Wang, H.F. and M.P. Anderson. 1982. Introduction to groundwater modeling: finite difference and finite element methods. Freeman and Co., San Francisco, CA: 1.

Wang, H.F. and M.P. Anderson. 1982. ibid.: 2.

Wang, H.F. and M.P. Anderson. 1982. ibid.: 3.

Wang, H.F. and M.P. Anderson. 1982. ibid.: 5-17.

Wang, H.F. and M.P. Anderson. 1982. ibid.: 25.

Wang, H.F. and M.P. Anderson. 1982. ibid.: 45.

Wang, H.F. and M.P. Anderson. 1982. ibid.: 69.

Warner, J.W., W.E. Hanna, R.J. Landry, J.P. Wulforst, J.A. Neeley, R.L. Holmes, C.E. Rice. 1975. Soil survey of Suffolk County, New York. United States Department of

Agriculture Soil Conservation Service in cooperation with Cornell Agricultural Experiment Station, U.S. Government Printing Office: 1975 O-473-964: 101 p. + maps.

Warren, M., W. DeLaguna, and N. Lusczynski. 1968. Hydrology of Brookhaven National Laboratory and vicinity, Suffolk County, NY. *USGS Bulletin* 1156-C: 127 p.

Waterloo Hydrogeologic, Inc. 2006. Visual MODFLOW v. 4.2: user's guide by Waterloo Hydrogeologic, Inc., Waterloo, Canada: 654 p.

Weiss, L. 1954. Foraminifera and origin of the Gardiners Clay (Pleistocene), eastern Long Island, New York. *USGS Professional Paper* 254-G: 143-163.

Wexler, E.J. 1988a. 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.

Wexler, E.J. 1988b. Groundwater flow and solute transport at a municipal landfill site on Long Island, New York, part 3, simulation of solute transport. *USGS Water Resources Investigations Report* 86-4207: 46 p.

Wexler, E.J. and P.E. Maus. 1988. Groundwater flow and solute transport at a municipal landfill site on Long Island, New York, part 2, simulation of groundwater flow. *USGS Water Resources Investigations Report* 86-4106: 44 p.

Willmott, C.J. 1981. Validation of models. *Physical Geography* 2, no. 2: 184-194.

Winston, R.B. 2009. ModelMuse - A graphical user interface for MODFLOW-2005 and PHAST. *USGS Techniques and Methods* 6–A29: 52 p.

Winter, C. L. and D.M. Tartakovsky. 2008. A reduced complexity model for probabilistic risk assessment of groundwater contamination. *Water Resources Research*. 44(6). Available at http://onlinelibrary.wiley.com/doi/10.1029/2007WR006599/full.

Worthington, S. R. 1999. A comprehensive strategy for understanding flow in carbonate aquifers, Karst Modeling. Special Publication, 5, Charlottesville, VA: 30-37.

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

Ye, M., K.F. Pohlmann, J.B. Chapman, G.M. Pohll, and D.M. Reeves. 2010. A model-averaging method for assessing groundwater conceptual model uncertainty. *Groundwater* 48, no. 5: 716-728.

Ye, M., S.P. Neuman, P.D. Meyer and K. Pohlmann. 2005. Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff. *Water Resources Research* 41, no. 12: W12429. doi:10.1029/2005WR004260.

Zadeh, L. 1988. Fuzzy Logic. *Computer* 21, no. 4: 83-93.

Zheng, C. 2010. MT3DMS v.5.3: A modular three-dimensional multispecies transport model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems - supplemental user's guide. Available at http://hydro.geo.ua.edu/mt3d/mt3dms_v5_supplemental.pdf.

Zimmerman, H.J. 2000. An application-oriented view of modeling uncertainty. *European Journal of Operational Research* 122, no. 2: 190-198.

Appendix I: Descriptors of Groundwater Heads

Observation well	Groundwater head (in feet)					Data count
	maximum	minimum	mean	median	range	
S3529	30.2	22.32	25.68	25.54	7.88	446
S44581	28.43	22.52	24.26	23.87	5.91	42
S47747	21.17	14.94	17.66	17.65	6.23	173
S72127	19.28	13.99	15.69	15.32	5.29	78
S72131	22.51	17.69	19.47	19.21	4.82	78
S72136	23.34	18.04	19.95	19.72	5.3	66
S72138	22.52	17.65	18.93	18.35	4.87	30
S72149	13.77	13.59	13.68	13.68	0.18	2
S72150	15.59	15.41	15.50	15.5	0.18	2
S72153	17.15	13.6	15.29	15.01	3.55	17
S72159	20.19	20.09	20.14	20.14	0.10	2
S72160	7.31	7.11	7.21	7.21	0.20	2
S72161	6.80	6.39	6.60	6.61	0.41	4
S72162	4.67	3.96	4.24	4.14	0.71	5
S72163	3.28	2.95	3.12	3.11	0.33	2
S72164	3.19	2.87	3.03	3.03	0.32	2
S72165	3.34	2.79	3.07	3.06	0.55	2
S72167	4.72	2.81	3.58	3.32	1.91	5
S72168	3.92	1.91	2.73	2.38	2.01	5
S72169	4.96	3.35	4.01	3.80	1.61	5
S72170	7.03	2.89	4.04	3.32	4.14	5
S72171	3.43	1.97	2.57	2.26	1.46	5
S72172	2.69	1.58	2.20	2.20	1.11	5
S72173	2.84	2.67	2.76	2.75	0.17	2
S72815	31.25	24.22	26.24	25.97	7.03	54
S72816	31.85	24.96	27.39	27.28	6.89	93
S72818	23.38	19.84	21.48	21.47	3.54	72
S72819	23.48	19.79	21.33	21.31	3.69	58
S72820	23.61	19.71	21.42	21.39	3.90	61
S72821	23.35	18.04	19.92	19.72	5.31	74
S72822	23.18	18.03	19.93	19.73	5.15	65
S72823	21.96	20.96	21.54	21.7	1.00	3
S72824	21.49	20.9	21.20	21.19	0.59	2
S72827	18.64	15.02	16.51	16.41	3.62	78
S72828	18.67	15.05	16.56	16.49	3.62	75
S72829	15.67	12.22	13.09	12.78	3.45	44
S72833	23.81	17.64	19.45	19.16	6.17	63

S72834	24.67	18.94	21.02	20.81	5.73	81
S72836	25.34	19.68	21.68	21.35	5.66	72
S72837	25.35	19.68	21.77	21.54	5.67	75
S73750	28.41	21.06	23.41	23.40	7.35	108
S73751	27.53	20.70	23.07	22.77	6.83	88
S73752	27.53	20.72	23.18	22.99	6.81	99
S73753	27.71	20.41	22.62	22.41	7.30	90
S73754	27.62	20.41	22.72	22.46	7.21	89
S73755	27.54	20.39	22.74	22.62	7.15	92
S73756	26.65	20.16	22.41	22.44	6.49	98
S73757	26.65	20.14	22.33	22.21	6.51	84
S73758	26.65	20.14	22.43	22.33	6.51	118
S73759	26.64	20.12	22.36	22.29	6.52	94
S73760	26.53	20.41	22.65	22.38	6.12	147
S73761	26.54	20.41	22.62	22.28	6.13	100
S73763	26.57	20.37	22.70	22.66	6.20	110
S73764	26.91	20.76	23.01	23.00	6.15	103
S73765	26.91	20.76	22.90	22.62	6.15	74
S73766	26.89	20.74	22.92	22.69	6.15	77
S73767	27.63	21.09	23.36	23.37	6.54	108
S73768	27.64	21.09	23.32	23.13	6.55	86
S73943	24.73	20.28	22.07	21.99	4.45	61
S73944	24.76	20.31	22.03	21.88	4.45	50
S73945	24.69	18.94	20.98	20.69	5.75	59
S73946	23.62	18.69	20.76	20.50	4.93	78
S73947	23.58	18.67	20.65	20.50	4.91	63
S73953	21.47	16.54	18.23	18.00	4.93	72
S73954	21.51	16.58	18.28	18.08	4.93	61
S76380	15.07	14.55	14.81	14.82	0.52	10
S76381	15.16	14.61	14.83	14.78	0.55	9
S76383	15.09	14.63	14.84	14.82	0.46	6
S76384	15.12	14.59	14.88	14.92	0.53	6
S76385	15.20	14.65	14.89	14.89	0.55	6
S76400	24.65	19.09	20.94	20.65	5.56	58
S76401	24.65	19.08	20.93	20.64	5.57	58
S95303	25.39	19.68	21.69	21.41	5.71	69
S95304	25.33	19.67	21.64	21.36	5.66	69
S95305	25.31	19.50	21.45	21.31	5.81	65
S95306	25.11	19.50	21.41	21.14	5.61	66
S95309	22.23	17.69	19.32	19.17	4.54	55
S95310	18.80	15.15	16.51	16.41	3.65	67

S95312	20.16	15.59	17.15	16.88	4.57	64
S95318	26.99	20.81	22.96	22.63	6.18	72
S95319	26.38	20.37	22.42	22.13	6.01	69
S95320	26.37	20.4	22.37	22.06	5.97	57
S95323	13.26	10.23	11.01	10.94	3.03	67
S95324	9.97	7.35	7.93	7.84	2.62	62
S95325	9.95	7.36	7.93	7.79	2.59	52
S96201	9.82	7.51	8.13	8.07	2.31	66
S96202	15.41	11.93	12.91	12.82	3.48	66
S98434	9.86	7.55	8.19	8.13	2.31	68
S98435	10.00	7.37	7.98	7.94	2.63	54
S98436	13.21	10.21	10.98	10.89	3.00	65
S98437	13.28	10.23	11.00	10.90	3.05	65
S98438	14.59	11.25	12.19	12.12	3.34	66
S98439	14.19	11.37	12.27	12.26	2.82	54
S98440	14.64	11.26	12.21	12.17	3.38	54
S98441	15.19	11.58	12.67	12.62	3.61	54
S98442	15.18	11.60	12.67	12.58	3.58	65
103140	26.53	20.46	22.40	22.13	6.07	65
103141	24.94	20.95	22.36	21.78	3.99	38
S72151M	17.23	13.52	14.56	14.48	3.71	73
S72812M	29.81	21.95	25.31	25.05	7.86	226
S72813M	26.49	20.51	22.80	22.62	5.98	234
Methane 5	28.40	22.39	23.80	23.69	6.01	43
MRF 4	29.22	22.92	24.91	24.71	6.30	53
MRF-1	26.47	20.41	22.44	22.23	6.06	67
MRF-2	24.97	20.03	21.54	20.72	4.94	33
MRF-3	26.82	20.63	22.68	22.44	6.19	68
MW10-D	28.83	21.96	23.98	23.85	6.87	126
MW10-I	28.78	21.89	23.84	23.66	6.89	108
MW10-S	28.77	21.82	23.92	23.78	6.95	127
MW11-M	27.97	21.66	23.68	23.57	6.31	123
MW12-I	28.02	25.01	26.29	26.08	3.01	16
MW1-S	29.31	22.05	24.41	24.20	7.26	114
MW2-D	29.16	22.08	24.29	24.09	7.08	141
MW2-S	29.15	22.04	24.28	24.08	7.11	143
MW3-S	28.26	21.30	23.56	23.44	6.96	133
MW4-D	27.25	21.00	23.13	23.03	6.25	135
MW4-S	27.26	20.98	23.13	22.98	6.28	149
MW5-D	33.49	25.47	28.06	28.00	8.02	129
MW5-I	33.53	25.47	28.04	28.01	8.06	115

MW5-S	33.53	25.5	28.07	28.00	8.03	142
MW6-D	32.10	24.15	26.88	26.67	7.95	112
MW6-S	32.14	24.16	26.83	26.56	7.98	124
MW7-S	30.57	23.38	25.55	25.53	7.19	123
MW8-D	29.96	22.11	25.00	24.89	7.85	122
MW8-I	29.93	22.95	24.91	24.76	6.98	105
MW8-S	29.88	22.87	24.95	24.88	7.01	124
MW9-S	29.43	22.26	24.51	24.42	7.17	124
PZ-1	30.18	23.28	25.77	25.62	6.90	108
PZ-2	28.25	21.85	23.84	23.70	6.40	119
PZ-3	29.76	23.3	25.32	25.22	6.46	102
PZ-4	29.38	22.89	24.85	24.74	6.49	105
PZ-5	26.88	22.44	24.34	24.24	4.44	101
PZ-6	28.32	21.96	23.89	23.74	6.36	104

Appendix II: Groundwater Heads - Observed and Simulated

#	Well	Observed Values (feet)			Simulated Values (feet)		
		min	median	max	min	median	max
1	103140	20.46	22.13	26.53	14.07	21.79	27.80
2	103141	20.95	21.78	24.94	14.16	21.89	27.39
3	S3529	22.32	25.55	30.20	16.22	23.97	30.19
4	S44581	22.52	23.88	28.43	10.43	22.05	28.15
5	S47747	14.94	17.65	21.17	9.76	15.15	21.65
6	S72127	13.99	15.32	19.28	7.86	14.32	20.41
7	S72131	17.69	19.22	22.51	11.07	17.49	23.49
8	S72136	18.04	19.73	23.34	11.53	17.73	23.93
9	S72138	17.65	18.36	22.52	11.20	17.29	23.54
10	S72149	13.59	13.68	13.77	6.73	10.52	17.56
11	S72150	15.41	15.50	15.59	7.93	12.42	19.34
12	S72151M	13.52	14.48	17.23	8.02	13.69	19.95
13	S72153	13.60	15.01	17.15	5.43	11.84	19.38
14	S72159	20.09	20.14	20.19	4.27	16.04	22.34
15	S72160	7.11	7.21	7.31	2.64	6.27	13.11
16	S72161	6.39	6.61	6.80	0.67	5.25	12.01
17	S72162	3.96	4.14	4.67	0.58	3.43	11.42
18	S72163	2.95	3.12	3.28	0.59	1.37	6.23
19	S72164	2.87	3.03	3.19	0.45	1.28	4.52
20	S72165	2.79	3.07	3.34	0.63	1.80	5.42
21	S72167	2.81	3.32	4.72	1.26	3.18	7.93
22	S72168	1.91	2.38	3.92	0.65	2.55	5.90
23	S72169	3.35	3.80	4.96	2.14	4.29	10.03
24	S72170	2.89	3.32	7.03	1.86	4.30	10.04
25	S72171	1.97	2.26	3.43	0.67	2.91	7.14
26	S72172	1.58	2.20	2.69	0.26	2.01	18.06
27	S72173	2.67	2.76	2.84	0.41	2.56	15.70
28	S72812M	21.95	25.05	29.81	2.92	23.91	29.62
29	S72813M	20.51	22.63	26.49	13.79	21.37	26.77
30	S72815	24.22	25.98	31.25	13.57	26.20	32.74
31	S72816	24.96	27.28	31.85	12.16	27.18	33.65
32	S72818	19.84	21.48	23.38	12.16	19.31	25.26
33	S72819	19.79	21.32	23.48	11.10	19.32	25.26
34	S72820	19.71	21.39	23.61	11.11	19.31	25.25
35	S72821	18.04	19.73	23.35	9.53	17.62	23.84
36	S72822	18.03	19.73	23.18	9.79	17.73	23.94
37	S72823	20.96	21.70	21.96	7.94	17.29	23.58

38	S72824	20.90	21.20	21.49	9.52	17.32	23.58
39	S72827	15.02	16.41	18.64	8.82	13.63	20.51
40	S72828	15.05	16.49	18.67	9.03	13.97	20.76
41	S72829	12.22	12.79	15.67	6.88	11.17	17.72
42	S72833	17.64	19.16	23.81	11.07	17.50	23.50
43	S72834	18.94	20.81	24.67	12.33	19.18	25.04
44	S72836	19.68	21.36	25.34	13.03	20.45	26.06
45	S72837	19.68	21.54	25.35	13.03	20.45	26.05
46	S73750	21.06	23.41	28.41	14.05	21.47	27.61
47	S73751	20.70	22.78	27.53	14.05	21.46	27.54
48	S73752	20.72	22.99	27.53	14.05	21.44	27.42
49	S73753	20.41	22.41	27.71	13.70	20.99	26.85
50	S73754	20.41	22.46	27.62	13.64	20.95	26.81
51	S73755	20.39	22.62	27.54	13.64	20.94	26.72
52	S73756	20.16	22.45	26.65	13.38	20.67	26.29
53	S73757	20.14	22.22	26.65	13.38	20.67	26.30
54	S73758	20.14	22.34	26.65	13.38	20.67	26.30
55	S73759	20.12	22.30	26.64	13.38	20.67	26.29
56	S73760	20.41	22.38	26.53	13.78	21.32	27.13
57	S73761	20.41	22.29	26.54	13.70	21.17	26.93
58	S73763	20.37	22.67	26.57	13.78	21.31	27.13
59	S73764	20.76	23.00	26.91	13.42	21.74	27.71
60	S73765	20.76	22.62	26.91	12.91	21.74	27.70
61	S73766	20.74	22.69	26.89	12.94	21.74	27.69
62	S73767	21.09	23.38	27.63	10.84	22.11	28.29
63	S73768	21.09	23.14	27.64	10.84	22.09	28.28
64	S73943	20.28	21.99	24.73	8.27	20.36	26.19
65	S73944	20.31	21.88	24.76	8.28	20.35	26.18
66	S73945	18.94	20.69	24.69	8.28	19.17	25.04
67	S73946	18.69	20.51	23.62	8.28	18.48	24.54
68	S73947	18.67	20.50	23.58	8.28	18.48	24.54
69	S73953	16.54	18.00	21.47	10.08	15.53	22.08
70	S73954	16.58	18.08	21.51	10.08	15.53	22.08
71	S76380	14.55	14.83	15.07	7.57	11.83	18.85
72	S76381	14.61	14.78	15.16	7.57	11.82	18.85
73	S76383	14.63	14.83	15.09	7.57	11.82	18.85
74	S76384	14.59	14.92	15.12	7.57	11.82	18.85
75	S76385	14.65	14.89	15.20	7.57	11.82	18.85
76	S76400	19.09	20.66	24.65	9.53	19.16	25.04
77	S76401	19.08	20.65	24.65	10.43	19.17	25.03
78	S95303	19.68	21.41	25.39	13.03	20.45	26.04

79	S95304	19.67	21.36	25.33	13.03	20.45	26.04
80	S95305	19.50	21.31	25.31	12.64	19.76	25.41
81	S95306	19.50	21.15	25.11	6.94	19.75	25.41
82	S95309	17.69	19.17	22.23	5.15	17.49	23.49
83	S95310	15.15	16.41	18.80	5.15	14.33	21.64
84	S95312	15.59	16.88	20.16	4.66	15.20	21.22
85	S95318	20.81	22.63	26.99	4.66	22.11	28.47
86	S95319	20.37	22.13	26.38	4.63	21.46	27.53
87	S95320	20.40	22.06	26.37	5.15	21.45	27.51
88	S95323	10.23	10.94	13.26	6.18	9.53	16.75
89	S95324	7.35	7.85	9.97	4.53	6.78	14.28
90	S95325	7.36	7.80	9.95	4.53	6.76	14.28
91	S96201	7.51	8.08	9.82	4.50	6.11	14.84
92	S96202	11.93	12.83	15.41	4.51	9.37	18.83
93	S98434	7.55	8.13	9.86	4.48	6.09	14.83
94	S98435	7.37	7.95	10.00	4.53	6.77	14.28
95	S98436	10.21	10.89	13.21	6.18	9.55	19.04
96	S98437	10.23	10.90	13.28	6.18	9.55	16.74
97	S98438	11.25	12.12	14.59	6.89	10.68	17.90
98	S98439	11.37	12.27	14.19	6.89	10.69	17.90
99	S98440	11.26	12.17	14.64	6.89	10.69	19.38
100	S98441	11.58	12.62	15.19	6.88	10.32	19.38
101	S98442	11.60	12.58	15.18	6.88	10.32	19.35
102	Methane 5	22.39	23.69	28.40	10.70	21.48	27.61
103	MRF-1	20.41	22.23	26.47	13.61	20.80	26.52
104	MRF-2	20.03	20.72	24.97	13.37	20.47	26.25
105	MRF-3	20.63	22.45	26.82	13.76	21.02	26.88
106	MRF 4	22.92	24.71	29.22	15.31	22.88	28.98
107	MW10-D	21.96	23.86	28.83	15.65	25.81	32.29
108	MW10-I	21.89	23.66	28.78	15.64	25.81	32.29
109	MW10-S	21.82	23.78	28.77	16.35	25.81	32.31
110	MW11-M	21.66	23.57	27.97	15.55	23.78	29.64
111	MW12-I	25.01	26.09	28.02	16.43	27.05	33.50
112	MW1-S	22.05	24.20	29.31	15.61	24.16	31.08
113	MW2-D	22.08	24.09	29.16	15.43	23.78	30.44
114	MW2-S	22.04	24.08	29.15	15.41	23.75	30.47
115	MW3-S	21.30	23.44	28.26	14.66	22.84	29.48
116	MW4-D	21.00	23.03	27.25	14.30	22.26	28.54
117	MW4-S	20.98	22.98	27.26	14.30	22.28	28.65
118	MW5-D	25.47	28.00	33.49	17.50	28.54	35.13
119	MW5-I	25.47	28.01	33.53	17.25	28.54	35.13

120	MW5-S	25.50	28.00	33.53	18.83	28.56	35.15
121	MW6-D	24.15	26.67	32.10	18.30	27.02	33.71
122	MW6-S	24.16	26.57	32.14	17.81	27.04	33.74
123	MW7-S	23.38	25.53	30.57	17.18	25.36	31.85
124	MW8-D	22.11	24.89	29.96	16.58	24.74	31.16
125	MW8-I	22.95	24.76	29.93	16.58	24.73	31.16
126	MW8-S	22.87	24.89	29.88	16.58	24.74	31.18
127	MW9-S	22.26	24.43	29.43	16.07	24.27	30.71
128	PZ-1	23.28	25.62	30.18	16.07	25.74	31.76
129	PZ-2	21.85	23.70	28.25	15.31	23.35	29.01
130	PZ-3	23.30	25.23	29.76	17.14	25.30	31.79
131	PZ-4	22.89	24.74	29.38	16.61	24.77	31.21
132	PZ-5	22.44	24.24	26.88	15.93	24.10	30.51
133	PZ-6	21.96	23.75	28.32	15.45	23.59	30.02

Appendix III: Model Area Metric Values (A*)

#	Models	Model Area metric (A*) (feet)	min	max	median	range
1	178	1.257	0.330	6.690	1.060	6.360
2	265	1.441	0.530	4.190	1.300	3.660
3	200	1.666	0.230	6.500	1.610	6.270
4	245	1.717	0.390	2.870	1.740	2.480
5	177	1.738	0.540	2.950	1.700	2.410
6	204	1.740	0.260	6.390	1.710	6.130
7	216	1.745	0.260	6.390	1.710	6.130
8	243	1.814	0.530	3.680	1.770	3.150
9	212	1.826	0.470	3.230	1.810	2.760
10	141	1.830	0.540	2.970	1.850	2.430
11	189	1.832	0.470	3.170	1.850	2.700
12	118	1.838	0.540	3.110	1.850	2.570
13	135	1.839	0.550	3.370	1.860	2.820
14	129	1.849	0.540	3.110	1.850	2.570
15	192	1.861	0.270	6.380	1.810	6.110
16	210	1.864	0.330	6.640	1.860	6.310
17	50	1.869	0.440	6.820	1.680	6.380
18	191	1.871	0.350	3.350	1.890	3.000
19	114	1.873	0.740	2.980	1.800	2.240
20	197	1.877	0.370	3.200	1.870	2.830
21	207	1.882	0.490	3.670	1.920	3.180
22	254	1.892	0.540	3.720	1.880	3.180
23	13	1.893	0.580	3.070	1.890	2.490
24	209	1.895	0.360	3.190	1.910	2.830
25	87	1.896	0.540	3.130	1.900	2.590
26	1	1.896	0.570	3.060	1.890	2.490
27	183	1.897	0.570	3.270	1.860	2.700
28	147	1.900	0.570	3.550	1.910	2.980
29	25	1.902	0.570	3.160	1.880	2.590
30	231	1.908	0.520	3.660	1.880	3.140
31	159	1.909	0.560	3.130	1.940	2.570
32	171	1.910	0.550	3.070	1.940	2.520
33	123	1.919	0.550	3.390	1.960	2.840
34	195	1.919	0.500	3.740	1.940	3.240
35	149	1.923	0.440	3.290	1.940	2.850
36	143	1.926	0.500	3.570	1.940	3.070

37	257	1.927	0.360	6.730	1.900	6.370
38	37	1.934	0.570	3.210	1.910	2.640
39	71	1.935	0.530	3.290	1.910	2.760
40	179	1.935	0.490	4.210	1.970	3.720
41	138	1.939	0.380	6.780	1.890	6.400
42	246	1.940	0.340	6.700	1.960	6.360
43	130	1.943	0.550	6.760	1.930	6.210
44	49	1.945	0.570	3.190	1.950	2.620
45	233	1.946	0.390	3.350	1.950	2.960
46	79	1.947	0.530	3.270	1.920	2.740
47	180	1.950	0.430	6.620	1.940	6.190
48	83	1.952	0.740	3.140	1.930	2.400
49	27	1.952	0.570	3.450	2.000	2.880
50	33	1.953	0.740	3.260	1.960	2.520
51	7	1.955	0.730	3.610	1.910	2.880
52	166	1.956	0.550	6.780	1.890	6.230
53	120	1.957	0.420	6.530	1.960	6.110
54	126	1.959	0.380	6.780	1.910	6.400
55	154	1.959	0.540	6.790	1.890	6.250
56	214	1.961	0.390	6.610	1.940	6.220
57	43	1.9631	0.730	3.320	1.920	2.590
58	186	1.963	0.400	6.860	1.940	6.460
59	80	1.964	0.550	6.990	1.920	6.440
60	132	1.966	0.450	6.570	1.980	6.120
61	206	1.968	0.220	6.160	1.970	5.940
62	84	1.976	0.740	7.130	1.950	6.390
63	34	1.979	0.690	7.140	1.970	6.450
64	260	1.979	0.520	3.880	2.000	3.360
65	72	1.980	0.550	7.010	1.890	6.460
66	137	1.985	0.430	3.540	2.040	3.110
67	247	1.988	0.230	4.590	1.970	4.360
68	4	1.991	0.570	6.810	1.930	6.240
69	202	1.992	0.390	6.620	2.000	6.230
70	31	1.993	0.730	3.750	1.960	3.020
71	75	1.996	0.740	7.150	1.940	6.410
72	104	1.999	0.540	7.040	1.950	6.500
73	10	2.002	0.680	7.110	1.980	6.430
74	96	2.002	0.540	7.050	1.940	6.510
75	108	2.004	0.750	7.180	1.960	6.430

76	99	2.005	0.750	3.280	1.980	2.530
77	218	2.008	0.210	6.150	2.030	5.940
78	119	2.009	0.480	4.050	2.030	3.570
79	198	2.010	0.330	6.640	1.970	6.310
80	16	2.010	0.570	6.800	1.990	6.230
81	22	2.011	0.670	7.100	1.970	6.430
82	136	2.012	0.530	7.000	1.980	6.470
83	69	2.014	0.700	7.190	1.980	6.490
84	185	2.014	0.450	3.550	2.050	3.100
85	14	2.017	0.540	6.980	1.970	6.440
86	174	2.018	0.380	6.830	2.010	6.450
87	9	2.018	0.730	3.470	2.020	2.740
88	95	2.019	0.530	3.580	2.050	3.050
89	100	2.020	0.750	7.200	1.980	6.450
90	115	2.021	0.740	7.220	1.940	6.480
91	162	2.023	0.390	6.840	2.000	6.450
92	92	2.024	0.740	7.170	1.970	6.430
93	107	2.025	0.750	3.270	2.050	2.520
94	125	2.026	0.430	3.550	2.080	3.120
95	222	2.032	0.350	6.660	2.050	6.310
96	3	2.033	0.570	3.940	2.060	3.370
97	156	2.038	0.430	6.590	2.050	6.160
98	90	2.041	0.590	6.840	2.040	6.250
99	172	2.042	0.540	7.040	1.980	6.500
100	103	2.045	0.530	3.530	2.090	3.000
101	21	2.047	0.730	3.430	2.080	2.700
102	131	2.050	0.480	4.030	2.100	3.550
103	38	2.052	0.530	7.040	1.960	6.510
104	155	2.052	0.490	4.420	2.050	3.930
105	190	2.054	0.390	6.610	2.050	6.220
106	228	2.058	0.300	5.750	2.090	5.450
107	94	2.061	0.630	6.950	2.050	6.320
108	46	2.061	0.690	7.170	2.020	6.480
109	40	2.066	0.560	6.870	2.050	6.310
110	15	2.067	0.570	3.900	2.060	3.330
111	160	2.067	0.550	7.060	1.980	6.510
112	168	2.068	0.420	6.580	2.060	6.160
113	240	2.078	0.300	5.720	2.100	5.420
114	124	2.085	0.540	7.010	2.040	6.470

115	57	2.085	0.680	7.160	2.090	6.480
116	111	2.092	0.550	7.070	2.050	6.520
117	52	2.102	0.560	6.860	2.080	6.300
118	194	2.103	0.210	6.160	2.130	5.950
119	259	2.107	0.170	6.480	2.140	6.310
120	161	2.108	0.440	4.090	2.110	3.650
121	116	2.115	0.740	4.550	2.110	3.810
122	55	2.116	0.690	7.360	2.010	6.670
123	45	2.134	0.740	4.000	2.180	3.260
124	184	2.135	0.560	7.080	2.080	6.520
125	208	2.135	0.490	6.870	2.110	6.380
126	261	2.141	0.590	6.820	2.110	6.230
127	173	2.143	0.430	4.030	2.130	3.600
128	238	2.143	0.430	5.600	2.190	5.170
129	224	2.143	0.160	6.410	2.160	6.250
130	20	2.144	0.690	7.310	2.060	6.620
131	77	2.154	0.630	6.930	2.140	6.300
132	117	2.161	0.640	7.010	2.140	6.370
133	226	2.166	0.430	5.640	2.250	5.210
134	39	2.167	0.560	4.500	2.130	3.940
135	244	2.168	0.510	6.930	2.160	6.420
136	56	2.169	0.730	3.940	2.180	3.210
137	82	2.169	0.590	6.810	2.130	6.220
138	193	2.174	0.260	5.010	2.270	4.750
139	44	2.181	0.690	7.380	2.080	6.690
140	258	2.184	0.230	4.570	2.230	4.340
141	167	2.185	0.480	4.520	2.240	4.040
142	73	2.188	0.590	4.800	2.190	4.210
143	85	2.190	0.620	6.920	2.180	6.300
144	51	2.200	0.550	4.430	2.200	3.880
145	8	2.207	0.680	7.330	2.130	6.650
146	236	2.209	0.160	6.470	2.190	6.310
147	248	2.209	0.160	6.470	2.190	6.310
148	250	2.215	0.430	5.680	2.350	5.250
149	252	2.218	0.260	5.050	2.310	4.790
150	36	2.236	0.550	6.920	2.200	6.370
151	98	2.243	0.600	6.890	2.150	6.290
152	102	2.270	0.630	7.000	2.190	6.370
153	151	2.273	0.260	5.010	2.360	4.750

154	67	2.273	0.690	7.400	2.250	6.710
155	235	2.276	0.230	4.760	2.340	4.530
156	35	2.278	0.550	4.910	2.320	4.360
157	106	2.280	0.600	6.880	2.240	6.280
158	255	2.293	0.520	6.960	2.270	6.440
159	109	2.308	0.630	6.990	2.250	6.360
160	145	2.315	0.340	5.500	2.440	5.160
161	101	2.316	0.740	4.980	2.290	4.240
162	29	2.322	0.520	5.390	2.410	4.870
163	32	2.328	0.680	7.360	2.300	6.680
164	140	2.332	0.200	6.540	2.350	6.340
165	128	2.332	0.200	6.550	2.350	6.350
166	188	2.336	0.210	6.630	2.340	6.420
167	12	2.350	0.540	6.910	2.320	6.370
168	230	2.353	0.200	5.960	2.450	5.760
169	262	2.353	0.740	4.910	2.360	4.170
170	97	2.364	0.570	5.470	2.370	4.900
171	70	2.364	0.560	6.990	2.300	6.430
172	134	2.371	0.300	6.370	2.450	6.070
173	242	2.374	0.200	5.910	2.520	5.710
174	122	2.381	0.260	6.320	2.440	6.060
175	6	2.391	0.510	6.630	2.400	6.120
176	24	2.391	0.540	6.900	2.380	6.360
177	105	2.407	0.570	5.390	2.450	4.820
178	11	2.424	0.540	5.530	2.430	4.990
179	139	2.425	0.250	5.570	2.470	5.320
180	18	2.442	0.510	6.620	2.490	6.110
181	23	2.478	0.540	5.470	2.500	4.930
182	121	2.482	0.330	6.100	2.550	5.770
183	5	2.488	0.520	6.000	2.500	5.480
184	187	2.491	0.250	5.710	2.530	5.460
185	164	2.496	0.200	6.620	2.460	6.420
186	48	2.523	0.550	6.980	2.440	6.430
187	17	2.557	0.520	5.950	2.620	5.430
188	133	2.557	0.320	6.040	2.650	5.720
189	158	2.574	0.260	6.390	2.570	6.130
190	59	2.585	0.540	6.970	2.540	6.430
191	42	2.588	0.520	6.700	2.540	6.180
192	181	2.610	0.210	6.410	2.570	6.200

193	170	2.652	0.260	6.390	2.660	6.130
194	53	2.662	0.520	6.690	2.630	6.170
195	163	2.717	0.240	6.400	2.670	6.160
196	157	2.723	0.200	6.680	2.720	6.480
197	47	2.736	0.520	6.350	2.650	5.830
198	175	2.813	0.240	6.310	2.810	6.070
199	41	2.846	0.130	6.840	2.820	6.710
200	169	2.926	0.140	6.730	3.060	6.590
Maximum		**2.926**	**0.750**	**7.400**	**3.060**	**6.650**
Minimum		**1.257**	**0.130**	**2.870**	**1.060**	**2.740**
Mean		**2.111**	**0.486**	**5.558**	**2.104**	**5.072**
Standard Deviation		**0.244**	**0.162**	**1.509**	**0.262**	**1.347**
median		**2.046**	**0.530**	**6.370**	**2.050**	**5.845**
1st quartile		**1.954**	**0.38**	**3.93**	**1.94**	**3.312**
3rd quartile		**2.238**	**0.57**	**6.845**	**2.25**	**6.37**